Check for updates

Received: 27 July 2021 | Revised: 2 March 2022 | Accepted: 2 March 2022

DOI: 10.1002/sim.9383

RESEARCH ARTICLE

Statistics in Medicine WILEY

Mixed-type multivariate response regression with covariance estimation

Karl Oskar Ekvall[1,2] | **Aaron J. Molstad**[3]

[1]Division of Biostatistics, Institute of Environmental Medicine, Karolinska Institutet, Stockholm, Sweden

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

[3]Department of Statistics and Genetics Institute, University of Florida, Gainesville, Florida

Correspondence
Karl Oskar Ekvall, Division of Biostatistics, Institute of Environmental Medicine, Karolinska Institutet, Stockholm, Sweden.
Email: karl.oskar.ekvall@ki.se

Funding information
Austrian Science Fund, Grant/Award Number: P30690-N35; NSF, U.S., Grant/Award Number: DMS-2113589

Abstract

We propose a new method for multivariate response regression and covariance estimation when elements of the response vector are 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 in practice. 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.

KEYWORDS

covariance estimation, latent variable models, mixed-type response regression, multivariate regression

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 be interested in the dependence between various types of omic data in a particular genomic region (see Section 6.2), some binary (eg, presence of somatic mutations) and some continuous (eg, gene expression). Joint modeling of responses facilitates inference on their dependence, and it 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

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.
© 2022 The Authors. *Statistics in Medicine* published by John Wiley & Sons Ltd.

type. A case in point is the multivariate normal linear regression model which assumes a vector of responses $Y \in \mathbb{R}^r$ and vector of predictors $x \in \mathbb{R}^p$ satisfy

$$Y \sim \mathrm{N}_r(\mathcal{B}^\mathsf{T} x, \Sigma), \tag{1}$$

for some regression coefficient matrix $\mathcal{B} \in \mathbb{R}^{p\times r}$ and covariance matrix $\Sigma \in \mathbb{S}^r_{++}$, the set of $r \times r$ symmetric and positive definite matrices. Model (1) is fundamental in multivariate statistics and leads to relatively straightforward likelihood-based inference; the parameters are identifiable, have intuitive interpretations, and maximum likelihood estimates are computationally tractable even when r and p are relatively large (but smaller than the number of independent observations n). Here, our goal is the development of a likelihood-based method for mixed-type response regression that retains some of the useful properties of (1). We focus on settings where dependence between responses cannot be parsimoniously parameterized and interest is in inference on an unstructured $r \times r$ covariance matrix characterizing this dependence. Rather than a method for specific combinations of response distributions, of which there are many,[1-11] we pursue a unified method that can be adapted to many different combinations of response distributions with relative ease.

To develop our method, we assume there is a known link function $g : \mathbb{R}^r \to \mathbb{R}^r$ and latent vector $W \in \mathbb{R}^r$ such that

$$g\{\mathbb{E}(Y \mid W)\} = W \quad \text{and} \quad W \sim \mathrm{N}_r(\mathcal{B}^\mathsf{T} x, \Sigma). \tag{2}$$

That is, a latent vector satisfies the multivariate linear regression model and the observable responses are connected to that latent vector through their conditional mean. We further suppose that, conditionally on W, the elements of Y satisfy generalized linear models with linear predictors given by the elements of W (see detailed specification in Section 2). This leads to a likelihood based on n independent observations $\{(y_i, x_i) \in \mathbb{R}^r \times \mathbb{R}^p, i = 1, \dots, n\}$ which can be written

$$L_n(\mathcal{B}, \Sigma) = \prod_{i=1}^{n} \int_{\mathbb{R}^r} f(y_i \mid w_i)\ \phi(w_i; \mathcal{B}^\mathsf{T} x_i, \Sigma)\ \mathrm{d}w_i, \tag{3}$$

where $f(y_i \mid w_i)$ is the conditional density for $Y_i = [Y_{i1}, \dots, Y_{ir}]^\mathsf{T}$ given $W_i = [W_{i1}, \dots, W_{ir}]^\mathsf{T}$ and $\phi(\cdot; \mathcal{B}^\mathsf{T} x_i, \Sigma)$ is the multivariate normal density with mean $\mathcal{B}^\mathsf{T} x_i$ and covariance matrix Σ. Note (1) is a special case of (2) where $Y = W$ and g is the identity. Conversely, (3) can be obtained as the likelihood for a specific generalized linear latent and mixed model (GLLAMM),[12] a class of models which also includes many other models for mixed-type responses as special cases.[13-18] Example 1 illustrates our setting.

Example 1. Suppose $Y \in \mathbb{R}^{20}$ is a vector of 10 count variables $(Y_1, \dots, Y_{10})$ and 10 continuous variables $(Y_{11}, \dots, Y_{20})$ whose joint distribution is of interest. If there are no predictors, a possible version of (2) assumes $Y_j | W \sim \mathrm{Poi}(W_j)$ for $j = 1, \dots, 10$, $Y_j \mid W \sim \mathrm{N}(W_j, 1)$ for $j = 11, \dots, 20$, and $W \sim \mathrm{N}_{20}(\mathcal{B}^\mathsf{T}, \Sigma)$, where $\mathcal{B}^\mathsf{T} \in \mathbb{R}^{20}$ and $\Sigma \in \mathbb{S}^{20}_{++}$.

There are several challenges with inference based on (3). First, in general the integrals cannot be computed analytically, and numerical integration can be prohibitively time consuming even for $r \approx 10$. This is commonly addressed by, for example, Laplace approximations, Monte Carlo integration, or penalized quasi likelihood (PQL), which results from dropping terms in a Laplace approximation that are assumed to be of lower order.[19-23] Here, in the interest of developing a method operational for relatively large r, we will approximate the likelihood using a modified version of PQL (Section 3), which is known to be fast.[21] We will also discuss how the proposed method can be extended to settings where the likelihood is approximated by other means.

A second challenge is that, even if the integrals in (3) can be computed efficiently, maximum likelihood estimation of $(\mathcal{B}, \Sigma)$ requires optimization over $\mathbb{R}^{p\times r} \times \mathbb{S}^r_{++}$:

$$(\hat{\mathcal{B}}, \hat{\Sigma}) \in \underset{(\mathcal{B},\Sigma)\in\mathbb{R}^{p\times r}\times\mathbb{S}^r_{++}}{\arg\max}\ L_n(\mathcal{B}, \Sigma). \tag{4}$$

When r is small it may be possible to solve (4) using off-the-shelf optimizers (eg, the `ml` command in Stata or the optim function in R). For example, if $r = 2$, then (4) can be characterized as a constrained optimization problem over $(\mathcal{B}, [\Sigma_{11}, \Sigma_{12}, \Sigma_{22}]^\mathsf{T}) \in \mathbb{R}^{p\times r} \times \mathbb{R}^3$, the constraints being that $\Sigma_{11} > 0$, $\Sigma_{22} > 0$, and $\Sigma_{12}^2 < \Sigma_{11}\Sigma_{22}$. When r is even moderately large, however, such constraints become untenable. In software packages for mixed and latent variable models it is common to maximize the likelihood in $\mathcal{B}$ and the Cholesky root L such that $\Sigma = LL^\mathsf{T}$.[24,25] This ensures the estimate of Σ is positive semidefinite, but it does not ensure positive definiteness and it complicates more ambitious inference.

For example, implementing a likelihood-ratio test of whether some of the responses are independent requires solving (4) subject to the constraint that some off-diagonal elements of Σ are zero (Section 4). Similarly, for some response-types it is necessary to constrain diagonal elements of Σ to ensure identifiability (see Section 2), and sometimes it is desirable to assume diagonal elements of Σ corresponding to responses of the same type (eg, responses 1-10 or 11-20 in Example 1) are the same. These types of restrictions are nontrivial to accommodate when parameterizing in terms of L, but as we will see in Section 3, they are elegantly handled by the method we propose. Briefly, our method iteratively updates a PQL-like approximation of (3) and maximizes that approximation in $\mathcal{B}$ and Σ using block coordinate descent. The update for $\mathcal{B}$ is a least squares problem which can be solved in closed form and the update for Σ is solved using an accelerated projected gradient descent algorithm. The algorithm scales well in the dimension r and it natively supports restrictions of the form $\Sigma \in \mathbb{M}$ for sets $\mathbb{M} \subseteq \mathbb{S}^r_{++}$ onto which a projection can be computed.

In some settings it may be appropriate to, as is common in mixed models, restrict $\Sigma = ZDZ^\mathsf{T}$ for some design matrix $Z \in \mathbb{R}^{r\times d}$ and covariance matrix $D \in \mathbb{S}^d_{++}$ of $d < r$ random effects (see eg, Jiang[26]). Such structures may be suggested by subject-specific knowledge or the sampling design, or they may be motivated by necessity when n is small relative to r. Similarly, many methods based on the marginal moments, such as for example generalized estimating equations,[27] specify a covariance structure. In many applications, however, it is unclear whether there is any particular dependence structure. Indeed, it may be of primary interest to discover a structure using data, rather than imposing one a priori; our method enables such discoveries. Some methods based on the marginal moments also allow the specification of an unstructured correlation matrix, but for Y rather than W.[28,29] In general, however, such methods are inconsistent with (2) since the mean vector and correlation matrix of Y cannot be independently parameterized. Accordingly, it is often unclear whether the estimates are in the parameter set of any particular model for mixed-type responses. Nevertheless, methods based on the marginal moments can be useful in practice and they are a reasonable comparison to our method for some purposes.

Other advantages of the parameterization we consider are that (i) the off-diagonal elements of Σ affect the joint distribution of responses but not their marginal (univariate) distributions and (ii) responses are independent if the corresponding off-diagonal elements of Σ are zero. Thus, we can test the null hypothesis that some responses are independent by testing whether the corresponding off-diagonal elements of Σ are zero, without that null hypothesis implying restrictions on the marginal distributions. This is generally not possible in more parsimonious parameterizations of the form $\Sigma = ZDZ^\mathsf{T}$ since, in general, elements of D affect both marginal and joint distributions.

2 | MODEL

2.1 | Specification

We now specify our model in detail and generalize somewhat compared to the Introduction. Assume the elements of $Y = (Y_1, \dots, Y_r)$ are conditionally independent given W with conditional densities of the form

$$f(y_j \mid w) = \exp\left\{\frac{y_j w_j - c_j(w_j)}{\psi_j}\right\}, \tag{5}$$

where $\psi = [\psi_1, \dots, \psi_r]^\mathsf{T}$ is a vector of strictly positive dispersion parameters and c_j is a cumulant function for the distribution of $Y_j \mid W$. For example, $c_j'(W_j) = \mathbb{E}(Y_j \mid W)$ and $\psi_j c_j''(W_j) = \mathrm{var}(Y_j \mid W)$, with primes denoting derivatives. From these properties it follows that the link function g defined by $g\{\mathbb{E}(Y \mid W\} = W$ satisfies $g(W) = [g_1(W_1), \dots, g_r(W_r)]^\mathsf{T}$ with g_j real-valued and strictly increasing for $j = 1, \dots, r$. That is, the jth latent variable has a direct effect on the jth response but no other responses. Equation (5) specifies a (conditional) generalized linear model[30] (GLM), which when $\psi_j = 1$ specializes to a one-parameter exponential family distribution, as in Example 1. Here, we do not assume $\psi_j = 1$ for every j but we do assume the ψ_j are known.

Because it makes our development no more difficult, in what follows we consider a slightly more general version of (2) where

$$W \sim \mathrm{N}_r(X\beta, \Sigma), \tag{6}$$

for a nonstochastic design matrix $X \in \mathbb{R}^{r\times q}$ and $\beta \in \mathbb{R}^q$. The classical multivariate response regression setting in (2) which motivates our study is then 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 vec(·) the vectorization operator stacking the columns of its matrix argument. Unlike (2) where all responses have the same predictors, (6) allows distinct predictors for each response, as in seemingly unrelated regressions.[31]

With $\{(y_i, X_i) \in \mathbb{R}^r \times \mathbb{R}^{r\times p}, i = 1, \dots, n\}$, denoting independent realizations, the likelihood is

$$L_n(\beta, \Sigma) = |\Sigma|^{-n/2} \prod_{i=1}^{n} \int_{\mathbb{R}^r} \exp\left\{ \left(\sum_{j=1}^{r} \frac{y_{ij} w_{ij} - c_j(w_{ij})}{\psi_j} \right) - \frac{1}{2}(w_i - X_i\beta)^\top \Sigma^{-1}(y_i - X_i\beta) \right\} \mathrm{d}w_i. \tag{7}$$

To see the connection to GLLAMMs and other mixed models which are often written for all observations simultaneously, let $\mathcal{Y} = [Y_1^\top, \dots, Y_n^\top]^\top \in \mathbb{R}^{rn}$, $\mathcal{X} = [X_1^\top, \dots, X_n^\top]^\top \in \mathbb{R}^{rn\times q}$, and $\mathcal{E} \sim \mathrm{N}_{nr}(0, I_n \otimes \Sigma)$. Then (7) is the likelihood for a model which assumes the elements of $\mathcal{Y}$ follow conditionally independent GLMs given $\mathcal{E}$, with canonical link functions and linear predictors given by the elements of

$$\mathcal{W} = \mathcal{X}\beta + \mathcal{E};$$

that is, each of the rn responses has a linear predictor with its own random intercept, and the covariance matrix for those random intercepts is $I_n \otimes \Sigma$.

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,

$$\mathbb{E}(Y) = \mathbb{E}\{g^{-1}(W)\} \quad \text{and} \quad \mathrm{cov}(Y) = \mathrm{cov}\{g^{-1}(W)\} + \mathbb{E}\{\mathrm{cov}(Y \mid W)\}. \tag{8}$$

We make a number of observations based on (8): first, because $\mathrm{cov}(Y \mid W)$ is assumed diagonal, the covariance between responses is determined by $\mathrm{cov}\{g^{-1}(W)\}$. 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 $\mathrm{cov}(Y \mid W)$ are nonlinear and nonconstant in general, $\mathbb{E}(Y)$ and $\mathbb{E}\{\mathrm{cov}(Y \mid W)\}$ in general depend on both β and diagonal elements of Σ. Fourth, since $\mathrm{var}(Y_j)$ is increasing in ψ_j and $\mathrm{cov}\{g^{-1}(W)\}$ does not depend on ψ, $\mathrm{cor}(Y_j, Y_k)$ 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 (8), 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 $\mathrm{var}(Y_j \mid W) = \psi_j$ for $j = 1, 2$, and $\mathbb{E}(Y_j|W) = \exp(W_j)$ and $\mathrm{var}(Y_j \mid W) = \psi_j \exp(W_j)$ 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 $\psi_3 = \psi_4 = 1$, $Y_j|W \sim \mathrm{Poi}\{\exp(W_j)\}, j = 3, 4$. 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 (Supplementary Material) gives the following moments: $\mathbb{E}(Y_1) = X_1^\top\beta$, $\mathbb{E}(Y_3) = \exp(X_3^\top\beta + \Sigma_{33}/2)$, $\mathrm{var}(Y_1) = \psi_1 + \Sigma_{11}$, $\mathrm{var}(Y_3) = \exp(2X_3^\top\beta + \Sigma_{33})\{\exp(\Sigma_{33}) - 1 + \psi_3 \exp(-X_3^\top\beta - \Sigma_{33}/2)\}$, $\mathrm{cov}(Y_1, Y_2) = \Sigma_{21}$, $\mathrm{cov}(Y_1, Y_3) = \Sigma_{31}\exp(X_3^\top\beta + \Sigma_{33}/2)$, and $\mathrm{cov}(Y_3, Y_4) = \exp(X_3^\top\beta + X_4^\top\beta + \Sigma_{33}/2 + \Sigma_{44}/2)\{\exp(\Sigma_{43}) - 1\}$; the remaining entries of $\mathrm{cov}(Y)$ are the same as those given up to obvious changes in subscripts. Clearly, both $\mathbb{E}(Y)$ and $\mathrm{cov}(Y)$ depend on β and Σ, but regardless of type, the variance of Y_j is increasing in Σ_{jj}, the mean is increasing in $X_j^\top\beta$, 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, $\mathrm{cor}(Y_1, Y_3)^2 \leq \Sigma_{11}\Sigma_{33}/[(\psi_1 + \Sigma_{11})\{\exp(\Sigma_{33}) - 1 + \psi_3/\mathbb{E}(Y_3)\}] \leq \Sigma_{33}/\{\exp(\Sigma_{33}) - 1\}$. 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 $X_3^\top \beta = X_4^\top \beta$. Then $\text{cor}(Y_3, Y_4) = \{\exp(\Sigma_{43}) - 1\}\{\exp(\Sigma_{33}) - 1 + \psi_3/\mathbb{E}(Y_3)\}$. For small ψ_3, this correlation is approximately $(\exp(\Sigma_{43}) - 1)/(\exp(\Sigma_{33}) - 1)$, which for $|\Sigma_{43}| \leq \Sigma_{33}$ 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 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, nonconstant $g, h : \mathbb{R} \to \mathbb{R}$, 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 $(-\sqrt{\Sigma_{11}\Sigma_{22}}, \sqrt{\Sigma_{11}\Sigma_{22}})$, respectively.*

The proof is in the Supplementary Material. We illustrate the usefulness of this result in another example.

Example 3 (Normal and Bernoulli responses). Suppose $r = 2$ with $Y_1|W_1 \sim \text{N}(W_1, \psi_1)$ and Y_2 Bernoulli distributed with $\mathbb{E}(Y_2|W_2) = \text{logit}^{-1}(W_2) = 1/\{1 + \exp(-W_2)\}$. Suppose also for simplicity $W \sim \text{N}_2(\beta, \Sigma)$, $\beta \in \mathbb{R}2$. 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 one.[17,32] 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 $Y = [Y_1, Y_2]^\top$, properties of which we consider next.

Equation (8) implies $\text{cov}(Y_1, Y_2) = \int [\{t_1\phi_{\beta,\Sigma}(t)\}/\{1 + \exp(-t_2)\}]\,\mathrm{d}t - \beta_1\mathbb{E}(Y_2)$. The integral does not admit a closed form expression, but Lemma 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 1). To understand which values $\text{cov}(Y_1, Y_2)$ 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 $W_2 = \left(\sqrt{\Sigma_{22}}/\sqrt{\Sigma_{11}}\right) W_1$. Then one can show $\text{cor}(Y_1, Y_2) = 2\int \left[\left\{\sqrt{\Sigma_{11}}t\phi(t)\right\} / \left\{1 + \exp(-\sqrt{\Sigma_{22}}t)\right\}\right]\,\mathrm{d}t/\sqrt{\psi_1 + \Sigma_{11}}$. 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 link functions and distributions of $Y \mid W$, but the proof can be adapted to other settings. In the Supplementary Material, we state a result (Lemma B.4) which outlines conditions for identifiability more generally. Essentially, when $Y_j \mid W$ satisfies a GLM, it suffices that, for every j, the variance of Y_j is not a function of the mean of Y_j. If it is, as is the case when Y_j is Bernoulli distributed, restrictions on diagonal elements of Σ are needed for identifiability. The proof is in Supplementary Material.

Theorem 1. *Suppose $\{(y_i, X_i) \in \mathbb{R}^r \times \mathbb{R}^r \times q; i = 1, \ldots, n\}$ is an independent sample from our model and that (i) g_j, $j = 1, \ldots, r$, 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) $\mathcal{X}^\top\mathcal{X}$ is invertible, then distinct (β, Σ) correspond to distinct distributions for $\mathcal{Y}$.*

From a computational perspective, nonidentifiability 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 $w \mapsto \nabla c(w) = g^{-1}(w)$, where $c(w) = \sum_{j=1}^r c_j(w_j)$; this is essentially equivalent to linearization of the link function, which has been considered in other latent variable models.[19] More specifically, 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 $\text{cov}\{\mathbb{E}(Y \mid W)\} \approx \nabla^2 c(w)\Sigma\nabla^2 c(w)$. Approximating $\mathbb{E}\{\text{cov}(Y \mid W)\} = \mathbb{E}\{\text{diag}(\psi)\nabla^2 c(W)\} \approx \text{diag}(\psi)\nabla^2 c(w)$ leads to $\text{cov}(Y) \approx \text{diag}(\psi)\nabla^2 c(w) + \nabla^2 c(w)\Sigma\nabla^2 c(w) =: C(w, \Sigma)$. 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

$$Y_i \sim \text{N}_r\{m(w_i, \beta), C(w_i, \Sigma)\}, \tag{9}$$

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

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

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 (9). For further insight it is helpful to note that if w_i is set to the maximizer of the ith integrand in (7), then h_n is the same type of approximation of L_n as that used in PQL, specialized to our setting. The correspondence between linearization of the link function and PQL is detailed by Breslow.[22] The correspondence implies that, in addition to the moment-based motivation given here, h_n can be motivated as a Laplace approximation of L_n, with terms assumed to be of lower stochastic order ignored (see Breslow and Clayton[20] for details). Roughly speaking, the approximation will be better the closer the distribution of Y_i is to a normal. Because the latent variables are normal, we expect the distribution of Y_i to be close to normal if the distribution of $Y_i \mid W_i$ is. Now, to estimate variance parameters, conventional PQL makes further approximations which lead to a set of estimating equations. We proceed differently and avoid these approximations, both because they lack formal motivation (section 2.5 in Breslow and Clayton[20]) and because they do not in general lead to a simpler optimization problem for an unstructured and positive definite Σ. Thus, we will work directly with the approximate log-likelihood h_n and propose an algorithm that is substantially different from ones commonly used for mixed models.

With h_n as the starting point, 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. Motivated by the connection to Laplace approximations, we update the w_i by setting them to equal to the maximizers of the integrand in (7) with β and Σ fixed at their current iterates. To summarize, we propose a blockwise iterative algorithm whose $(k + 1)$th iterates are obtained using the updating equations

$$\left(\beta^{(k+1)}, \Sigma^{(k+1)}\right) = \underset{\beta, \Sigma}{\arg\min}\, h_n(\beta, \Sigma \mid w_1^{(k)}, \ldots, w_n^{(k)}), \tag{10}$$

$$\left(w_1^{(k+1)}, \ldots, w_n^{(k+1)}\right) = \underset{w_1, \ldots, w_n}{\arg\max} \sum_{i=1}^{n} \log f_{\beta^{(k+1)}, \Sigma^{(k+1)}}(y_i, w_i). \tag{11}$$

This algorithm can be run for a prespecified 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 (10) and (11). The final iterates of Σ and β are approximate MLEs and h_n evaluated at the final iterates of $w_1, \ldots, w_n$ is an approximate log-likelihood which we will use for approximate likelihood-based inference, including the construction of likelihood-ratio tests.

A formal statement of the proposed algorithm is in Algorithm 1.

3.2 | Updating β and Σ

To solve (10), we use a blockwise coordinate descent algorithm. Treating $w_1, \ldots, w_n$ as fixed throughout and ignoring the iterate superscript, this algorithm iterates between updating β and Σ. Specifically, the $(l + 1)$th iterates of the algorithm

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

for solving (10) can be expressed

$$\beta^{(l+1)} = \arg\min_\beta h_n\left(\beta, \Sigma^{(l)} \mid w_1, \dots, w_n\right), \tag{12}$$

$$\Sigma^{(l+1)} = \arg\min_\Sigma h_n\left(\beta^{(l+1)}, \Sigma \mid w_1, \dots, w_n\right). \tag{13}$$

Update (12) can be shown to be a weighted residual sum-of-squares with solution

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

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 nontrivial owing to nonconvexity 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 eg, Pinheiro and Bates[33]) 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 allows restrictions on the elements of Σ and ensures estimates are positive semidefinite, and can be constrained to be positive definite if desirable.

By picking an appropriate (convex) $\mathbb{M} \subseteq \mathbb{R}^r \times r$, (13) 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 nonconvexity and general constraints, we propose to solve this problem using a variation of the inertial proximal algorithm proposed by Ochs et al.[34] This is an accelerated projected gradient descent algorithm that can be used to minimize an objective function which is the sum of a nonconvex smooth function and convex nonsmooth function. In our case, h_n (as a function of Σ) is the nonconvex smooth function and the convex nonsmooth function is the function which equals ∞ if $\Sigma \notin \mathbb{M}$ and zero otherwise. This algorithm, like many popular accelerated first order algorithms, for example, FISTA,[35] uses "inertia" in the sense that the search point is informed by the direction of change from the previous iteration, which can lead to faster convergence.

To summarize briefly, our algorithm for (13) 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 $\gamma \in (0, 1)$, α is determined using backtracking line search (see algorithm 4 of Ochs et al[34]), 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 nonempty, closed, and convex (corollary 5.1.19 of Megginson[36]). 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 $\epsilon \ge 0$, where $\epsilon = 0$ ensures positive semidefiniteness and $\epsilon > 0$ positive definiteness. To compute projections onto $\mathbb{M}$ of this form, we implement Dykstra's alternating projection algorithm.[37] 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 Material. The algorithm is terminated when the objective function values converge.

3.3 | Updating the approximation points

We use a trust region algorithm for updating $w_i, i = 1, \dots, n$ (eg, chapter 4 of Nodecal and Wright[38]). 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 Material and the Hessian is, assuming Σ^{-1} is positive definite, for $i = 1, \dots, n$, $\nabla^2_{w_i} \log f_\theta(y_i, w_i) = -\nabla^2 c(w_i) - \Sigma^{-1}$. 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 $\underline{\Sigma} = \Sigma + \kappa I_r$ for some small $\kappa > 0$. Then the optimization problem for updating w_i is

$$\arg\min_{w \in \mathbb{R}^r} \left\{ -y_i^\top w + c(w) + \frac{1}{2}(w_i - X_i^\top \beta)^\top \underline{\Sigma}^{-1}(w_i - X_i\beta) + \tau \|w_i - X_i^\top \beta\|^2 \right\},$$

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 Supplementary 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

To make inferences about the parameters we use the approximate negative log-likelihood $h_n(\beta, \Sigma \mid w_1, \dots, w_n)$, where $w_1, \dots, w_n$ are held at the final iterates given by Algorithm 1. Focus is on testing hypotheses of the form $(\beta, \Sigma) \in \mathbb{H}_0$ vs $(\beta, \Sigma) \in \mathbb{H}_A$, where $\mathbb{H}_0$ and $\mathbb{H}_A$ partition the parameter set. If the null hypothesis constrains β only, we write for simplicity $\beta \in \mathbb{H}_0$, and similarly with $\Sigma \in \mathbb{H}_0$ if the null hypothesis constrains Σ only. We propose the test statistic $T_n = h_n(\tilde{\beta}, \tilde{\Sigma} \mid \tilde{w}_1, \dots, \tilde{w}_n) - h_n(\overline{\beta}, \overline{\Sigma} \mid \tilde{w}_1, \dots, \tilde{w}_n)$, where $(\tilde{\beta}, \tilde{\Sigma})$ and the approximation points $\tilde{w} = \{\tilde{w}_1, \dots, \tilde{w}_n\}$ are obtained by running Algorithm 1 with the restrictions implied by $\mathbb{H}_0$ and $(\overline{\beta}, \overline{\Sigma}) = \arg\min_{(\beta,\Sigma) \in \mathbb{H}_0 \cup \mathbb{H}_A} h_n(\beta, \Sigma | \tilde{w}_1, \dots, \tilde{w}_n)$. That is, $(\tilde{\beta}, \tilde{\Sigma})$ and $\tilde{w}$ are estimates and expansion points, respectively, from fitting the null model while $\overline{\beta}$ and $\overline{\Sigma}$ are obtained by maximizing the working likelihood from (9) 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 $(\overline{\beta}, \overline{\Sigma})$ 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 and $\mathbb{H}_0$ contains no boundary points of the parameter set, we expect T_n to be, under the null hypothesis, approximately chi-square distributed with degrees of freedom equal to the number of restrictions implied by $\mathbb{H}_0$.

A main motivation for our method is inference on the covariance matrix Σ. 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. For example, to test whether the first and second element of Y are independent, fitting the null model corresponds to setting

$$\mathbb{M} = \{\Sigma \in \mathbb{S}^r_+ : \Sigma_{12} = \Sigma_{21} = 0\},$$

assuming there are no other restrictions. Similarly, independence of more than two responses can be tested by including more off-diagonal restrictions in the definition of $\mathbb{M}$, and equality of variances for some responses can be tested by including restrictions such as $\Sigma_{11} = \Sigma_{22}$. In principle our method could also be used to test restrictions such as $\Sigma_{11} = 0$, corresponding to the first latent variable being constant. However, any null hypothesis which forces some eigenvalues of Σ to be zero corresponds to testing of boundary points, and then the likelihood ratio test-statistic has a different asymptotic distribution.[39-41]

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 5.5.

Algorithm 2. Hypothesis testing procedure for $(\beta, \Sigma) \in \mathbb{H}_0$ vs $(\beta, \Sigma) \in \mathbb{H}_A$

1. Given $\mathbb{H}_0$ and $\mathbb{H}_A$, initialize $(\beta^{(1)}, \Sigma^{(1)}) \in \mathbb{H}_0$.
2. For $k = 1, 2, \dots$ until convergence:
   (a) $w_i^{(k+1)} = \arg\min_{w \in \mathbb{R}^r} \left\{ -y_i^\top w + c(w) + \frac{1}{2}(w_i - X_i^\top \beta^{(k)})^\top (\Sigma^{(k)})^{-1}(w_i - X_i\beta^{(k)}) + \tau \|w_i - X_i^\top \beta^{(k)}\|^2 \right\}$ for $i = 1, \dots, n$
   (b) $(\beta^{(k+1)}, \Sigma^{(k+1)}) = \arg\min_{(\beta,\Sigma)\in\mathbb{H}_0} h_n(\beta, \Sigma \mid w_1^{(k+1)}, \dots, w_n^{(k+1)})$
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 $(\bar{\beta}, \bar{\Sigma}) = \arg\min_{(\beta,\Sigma)\in\mathbb{H}_0\cup\mathbb{H}_A} h_n(\beta, \Sigma \mid \tilde{w}_1, \dots, \tilde{w}_n)$
5. Return $T_n = h_n(\tilde{\beta}, \tilde{\Sigma} \mid \tilde{w}_1, \dots, \tilde{w}_n) - h_n(\bar{\beta}, \bar{\Sigma} \mid \tilde{w}_1, \dots, \tilde{w}_n)$.

5 | NUMERICAL EXPERIMENTS

5.1 | Overview

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. Specifically, when investigating prediction (Section 5.2) and estimation of β (Supplementary Material) we compare to separate GLMMs for the different response types. We chose this comparison because the GLMMs have correctly specified marginal distributions in our setting. In particular, the marginal moments $\mathbb{E}(Y_j)$ and $\text{var}(Y_j)$, $j = 1, \dots, r$, are correctly specified under the GLMMs. Thus, we expect the GLMMs to give reasonable estimates of the mean functions, and expect differences in predictive performance to our method to be indicative of the usefulness of joint modeling and estimating the off-diagonal elements of Σ. We also compare our method on prediction, estimation of β, and estimation of $\text{cor}(Y)$ to multivariate covariance generalized linear models (MCGLMs).[29] MCGLMs model the marginal moments of Y and, accordingly, provide estimates of the mean vector and covariance matrix of Y, but not of Σ. To further highlight the usefulness of modeling dependence, we include results for our method with Σ constrained to be diagonal, that is, with responses assumed independent. Whether constraining Σ to be diagonal or not, diagonal elements corresponding to Bernoulli responses are constrained to equal one when using our method.

5.2 | Prediction comparisons

We consider $r = 9$ response variables, three normally distributed, three Poisson, and three Bernoulli. When fitting separate GLMMs for the three-dimensional response vectors with elements of the same type, we consider two covariance structures that common software can fit: (i) $\Sigma_{jj} = \sigma_j^2$ and $\Sigma_{jk} = 0$ for $j \neq k$ or (ii) $\Sigma = \sigma^2 1_3 1_3^\top$, for some $\sigma^2 > 0$ and $1_3 = [1, 1, 1]^\top$. 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, respectively. There are many software packages for fitting GLMMs. We pick the `glmm`[23] package to fit the independent GLMMs and fit the clustered GLMMs using `lme4`.[25] Briefly, the former uses a Monte Carlo approximation of the likelihood and the latter uses adaptive Gaussian quadrature.

Predictions are formed by plugging estimates from the different methods into the expressions for $\mathbb{E}(Y_i) = \mathbb{E}\{\mathbb{E}(Y_i \mid W_i)\}$ in Section 2, and for simplicity we define prediction errors as the differences between those expectations and the observed responses, regardless of type. 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.

We next describe how data are generated in the simulations. The responses have different predictors and β is partitioned accordingly: $\beta = [\beta_1^\top, \dots, \beta_r^\top]^\top$, $\beta_j \in \mathbb{R}^{pj}$, and $q = \sum_{j=1}^r p_j$. We write $X_{i,j} \in \mathbb{R}^{pj}$ 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 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 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 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 all the structures, the GLMMs have correctly specified diagonal elements of Σ, but incorrectly specified off-diagonal elements in general.

For each structure of 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). We present relative squared out of sample 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, the proposed method performs best. When the covariance structure is nonsparse (eg, autoregressive or compound symmetric), the clustered GLMMs can outperform our method with the diagonal covariance matrix 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 specified covariance structure being a poor approximation to the

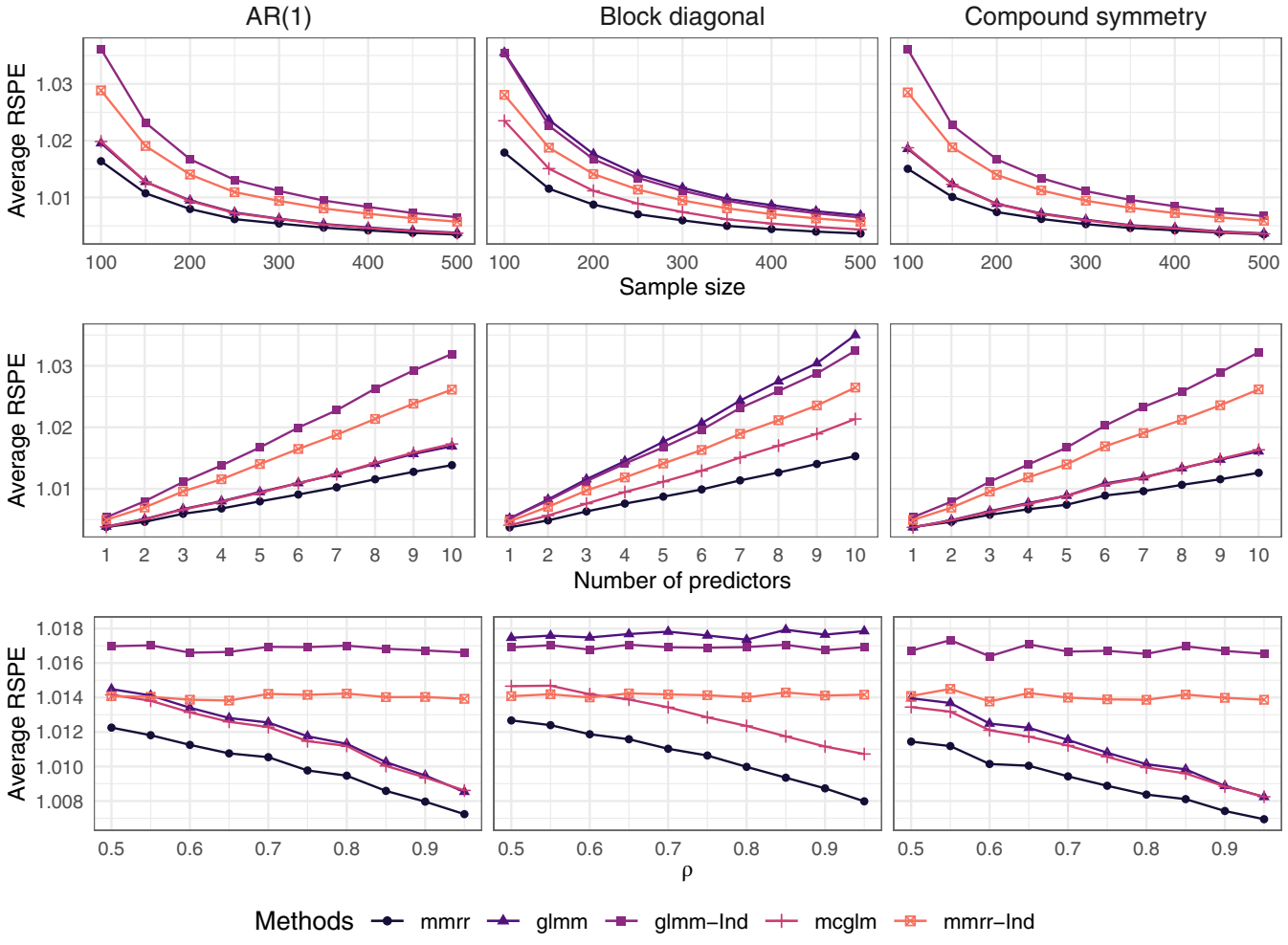

FIGURE 1 Average relative squared prediction errors. Top: $\rho = 0.9$ and $p_j = 5$ for $j = 1, \ldots, 9$. Middle: $n = 200$ and $\rho = 0.9$. Bottom: $n = 200$ and $p_j = 5$ for $j = 1, \ldots, 9$. glmm indicates clustered GLMMs, glmm-Ind GLMMs with diagonal covariance matrix, mcglm the method of Bonat and Jørgensen,[29] mmrr the proposed method, and mmrr-Ind the proposed method with diagonal covariance matrix

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. MCGLMs perform second-best in most settings, can perform similarly to clustered GLMMs when the true covariance structure is nonsparse, and can be outperformed by our method with diagonal covariance when dependence is weak. In summary, the results show the usefulness of joint modeling for prediction, and they illustrate the usefulness of the proposed method relative to ones based on marginal moments.

The Supplementary Material include results on mean squared errors for estimating β, for the same methods and settings used in Figure 1, and the results are qualitatively similar to those in Figure 1. The Supplementary Material also include a comparison to separate GLMs, effectively assuming independence between responses, which leads to substantially worse predictions than all methods considered here.

5.3 | 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 vs 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.

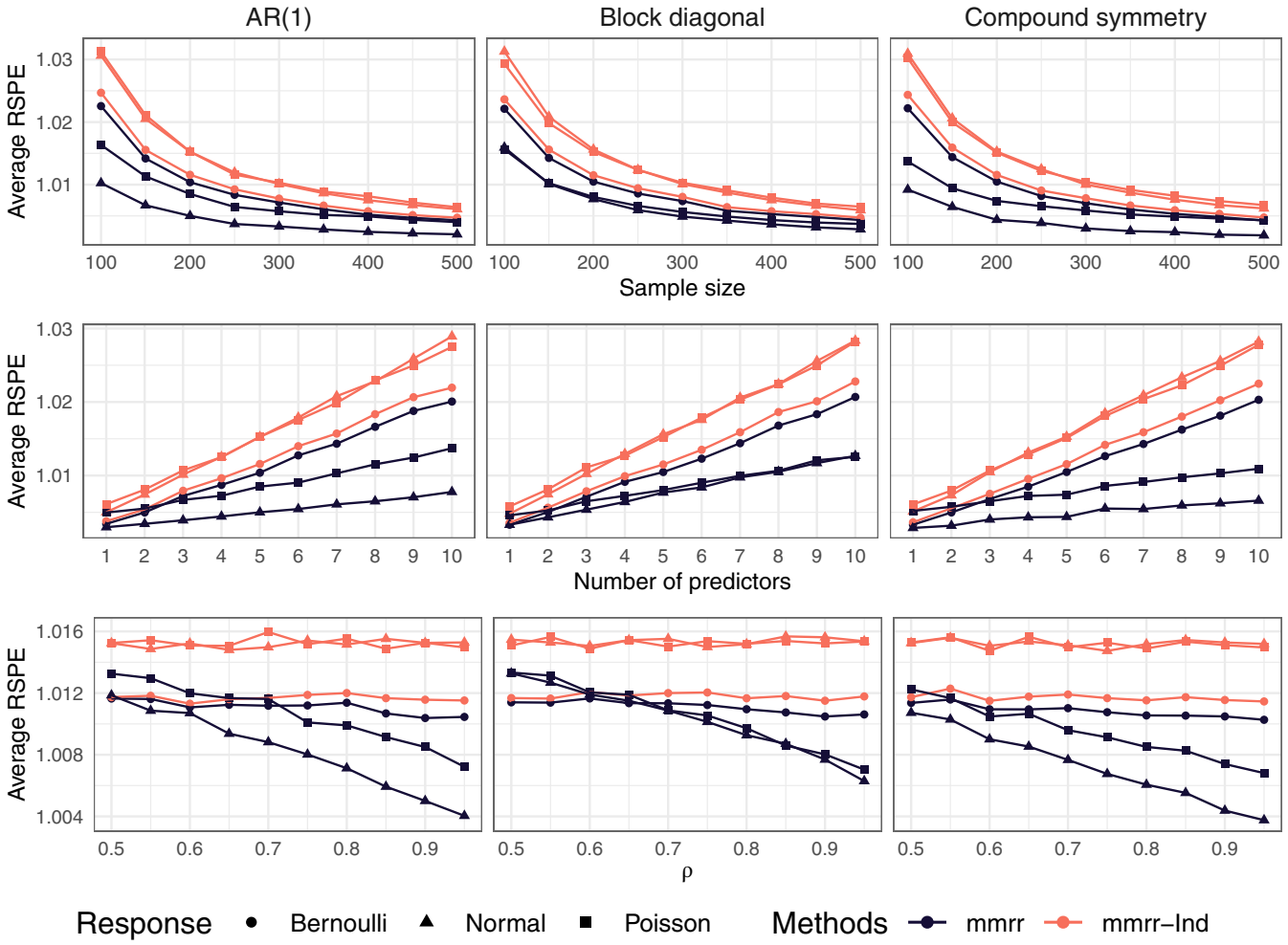

FIGURE 2 Average relative squared prediction errors. Top: $\rho = 0.9$ and $p_j = 5$ for $j = 1, \ldots, 9$. Middle: $n = 200$ and $\rho = 0.9$. Bottom $n = 200$ and $p_j = 5$ for $j = 1, \ldots, 9$. mmrr is the proposed method and mmrr-Ind the proposed method with diagonal covariance matrix

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 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. 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 Poisson.

In Section F.2 of the Supplementary Material, 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. We present similar results comparing our method and MCGLMs in the Supplementary Material.

5.4 | Covariance estimation

To investigate the usefulness of the proposed method for estimating covariances and correlations specifically, we consider a setting without predictors. That is, we consider (2) with $x = 1$, corresponding to an intercept only. Ideally we would compare to an established method for estimating Σ in our setting. Since none is available (to the best of our knowledge), we compare the estimates of $\Omega = \text{cov}(Y)$ and $\text{cor}(Y) = \text{diag}(\Omega)^{-1/2}\Omega\text{diag}(\Omega)^{-1/2}$ from our method to moment-based estimates. One is the empirical (or sample) covariance matrix $S = n^{-1}\sum_{i=1}^{n}(y_i - \bar{y})(y_i - \bar{y})^\top$, where $\bar{y} = n^{-1}\sum_{i=1}^{n} y_i$ is the sample mean. The corresponding empirical correlation matrix is $\text{diag}(S)^{-1/2}S\text{diag}(S)^{-1/2}$. MCGLMs also provide a moment-based estimate of $\text{cor}(Y)$, which we found to be indistinguishable from the empirical correlation matrix in the present setting without predictors.

We note the comparison to S, or other moment-based estimates of $\text{cov}(Y)$ not assuming (2), is not ideal because S cannot in general be mapped to an estimate of Σ. Formally the issue is that while Ω is injective as a function of Σ, which

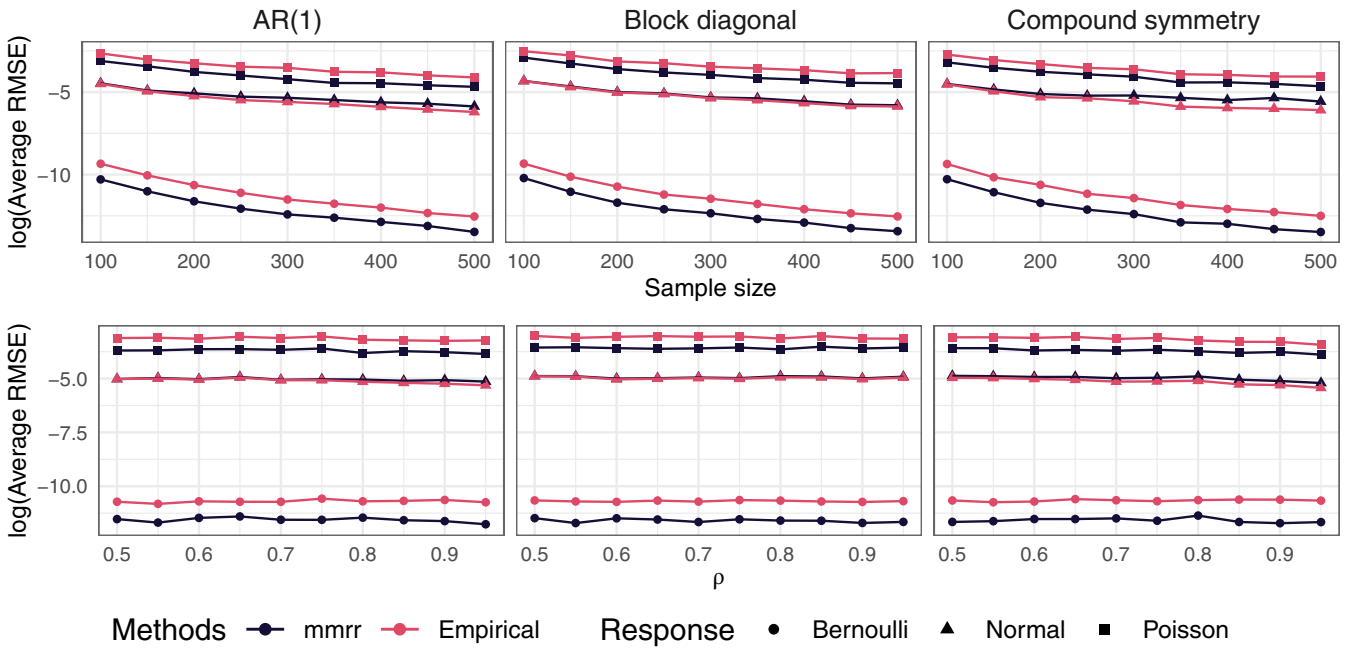

FIGURE 3 Average relative mean squared error for estimating the diagonals of $\text{cov}(Y)$ stratified by response type. (Top row) Results with $\rho = 0.9$. (Bottom row) Results with $n = 200$. mmrr is the proposed method and Empirical the sample covariance

follows from Theorem 1, that function is not onto $\mathbb{S}^r_+$. Put differently, not all realizations of S give an estimate consistent with (2). Nevertheless, S is (strongly) consistent for Ω when there are no predictors since the observations are then i.i.d., so we expect reasonable estimates of Ω.

The data generating model is as in Section 5.2, but with an intercept only. Figure 3 shows average relative mean square error for the diagonal entries of Ω by type. Specifically, we report the averages of: (normal) $\sum_{j=1}^{3}(\Omega_{jj}-\hat{\Omega}_{jj})^2/\sum_{j=1}^{3}\Omega_{jj}^2$, (Bernoulli) $\sum_{j=4}^{6}(\Omega_{jj}-\hat{\Omega}_{jj})^2/\sum_{j=4}^{6}\Omega_{jj}^2$, and (Poisson) $\sum_{j=7}^{9}(\Omega_{jj}-\hat{\Omega}_{jj})^2/\sum_{j=7}^{9}\Omega_{jj}^2$ where $\hat{\Omega}$ is an estimate of Ω with jth diagonal entry Ω_{jj}. In Figure 3, we see our method estimates the variances of the Poisson components more accurately than does the sample covariance, but the two perform similarly for normal and Bernoulli responses.

Figure 4 displays the average mean squared estimation error for $\text{diag}(\Omega)^{-1/2}\Omega\text{diag}(\Omega)^{-1/2}$, the correlation matrix of $(Y_1, \ldots, Y_9)^\top$. For smaller sample sizes, the proposed method performs better than MCGLMs. For larger sample sizes the differences diminish somewhat, which is not surprising given that sample correlations are consistent. As the correlation parameter ρ varies with $n = 200$, we see that the differences between the methods remain relatively constant, with the proposed one being better in every setting.

5.5 | 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 $\Sigma \in \mathbb{H}_0 = \mathbb{D}^r_{++}$ vs $\Sigma \in \mathbb{H}_A = \mathbb{S}^r_{++} \setminus \mathbb{H}_0$, and, assuming all responses have the same predictors as in (2), $\mathcal{B} \in \mathbb{H}_0 = \{\mathcal{B} \in \mathbb{R}^{p\times r} : \mathcal{B}_{kj} = 0,\ j = 1, \ldots, r\}$ vs $\mathcal{B} \in \mathbb{H}_A = \mathbb{R}^{p\times r} \setminus \mathbb{H}_0$, where $\mathcal{B}_{kj}$ denotes the kth predictor's effect on the jth response variable, that is, the (k,j)th element of $\mathcal{B}$. We set $k = 2$. Thus, the null hypothesis implies 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 5.3 but with $X_{i,1} = X_{i,2} = \cdots = X_{i,r}$ for all $i = 1, \ldots, n$ and $\mathcal{B} = [\beta_1, \ldots, \beta_r] \in \mathbb{R}^{p\times r}$. In the first setting, $n \in \{200,400, \ldots, 1000\}$ and $\tilde{\Sigma}_{jk} = \rho^{|j-k|}$, $\rho \in \{0.0, 0.05, \ldots, 0.4\}$. The top row of Figure 5 displays the proportion of rejections at the 0.05 significance level. When $\rho = 0$ (null hypothesis 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

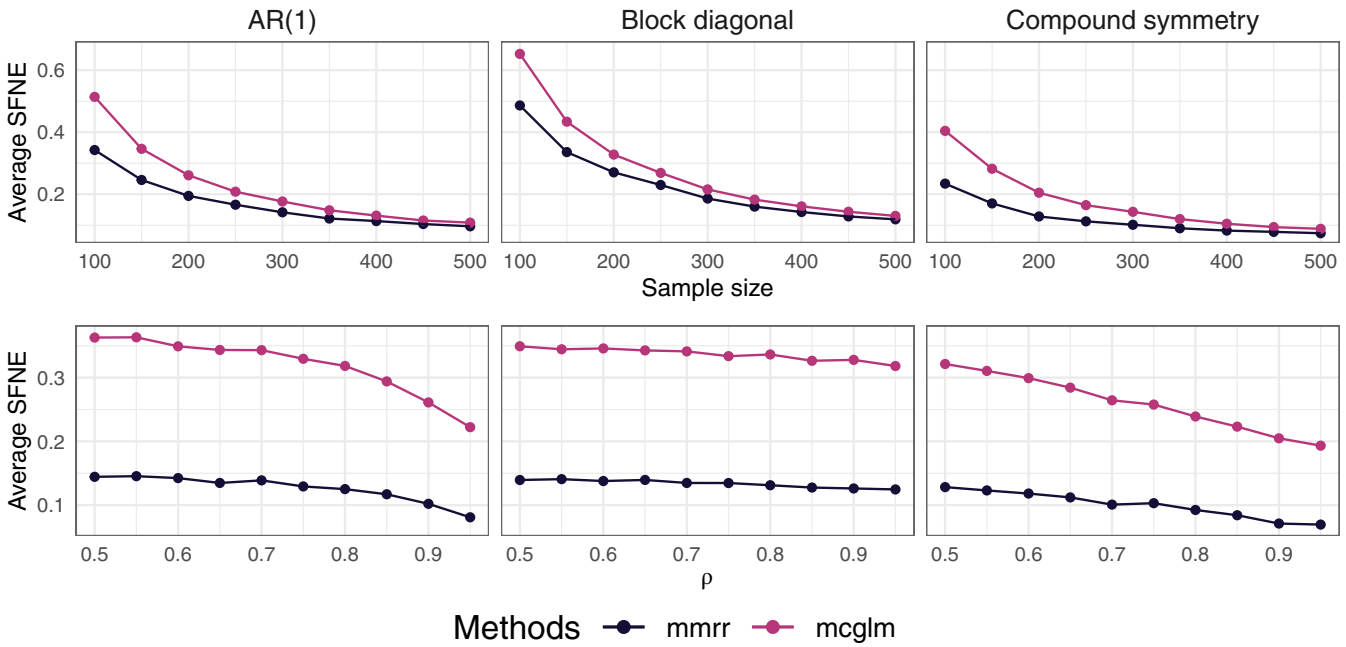

FIGURE 4 Average squared Frobenius norm error for estimating cor(Y). (Top row) Results with $\rho = 0.9$. (Bottom row) Results with $p = 1$ and $n = 200$. mmrr is the proposed method and mcglm the method of Bonat and Jørgensen[29]

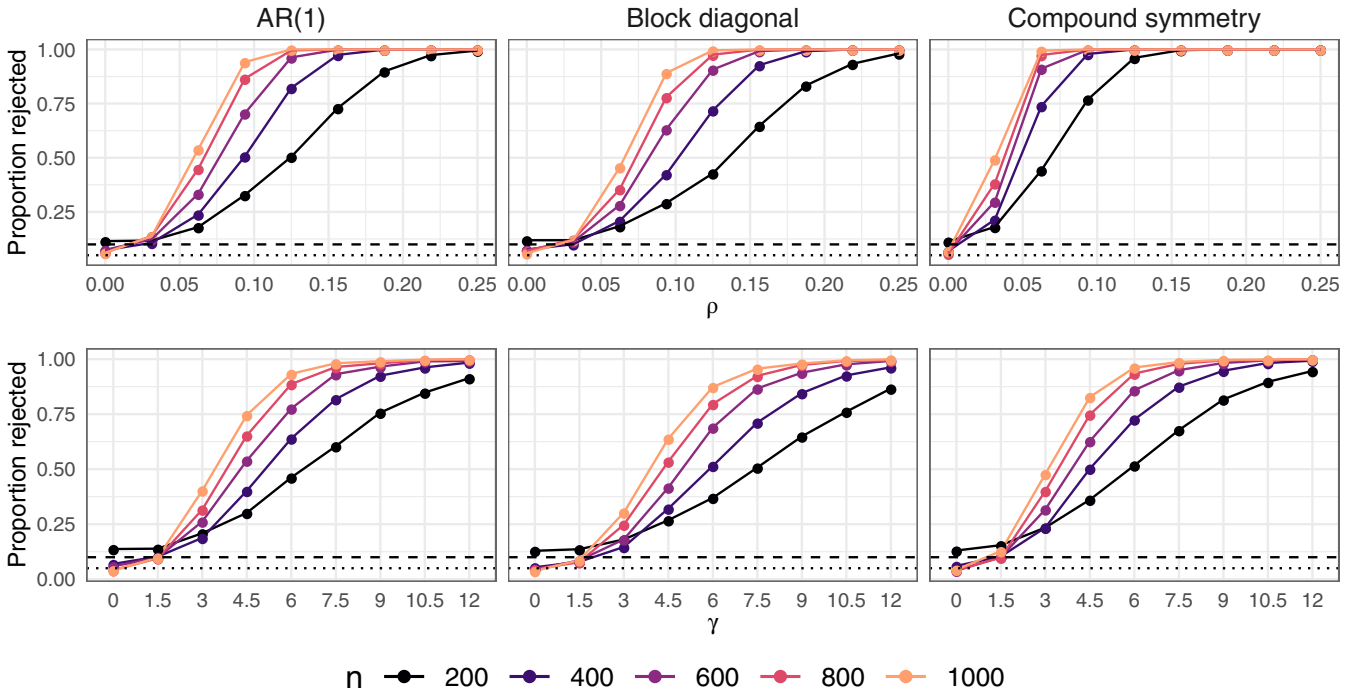

FIGURE 5 Top: Proportion of $\Sigma \in \mathbb{H}_0 = \mathbb{D}^r_{++}$ rejected at 0.05 level. Bottom: Proportion of $\mathcal{B} \in \mathbb{H}_0 = \{\mathcal{B} \in \mathbb{R}^{p\times r} : \mathcal{B}_{kj} = 0,\ j = 1, \dots, r\}$ rejected at the 0.05 level. Horizontal dashed and dotted black lines indicate 0.10 and 0.05

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 5.3, for $j = 1, \dots, r$ 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 5 shows that when $\gamma = 0$, so that $\mathcal{B}_{kj} = 0$ for all $j = 1, \dots, r$, 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 an 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

We consider a dataset collected on 333 women who were having difficulty becoming pregnant.[42] 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 are therefore 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. To illustrate how the proposed methods can be applied, we test the null hypothesis that Σ is diagonal and find evidence against it (P-value $< 10^{-16}$) using the test described in Section 5.5. 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, $\widehat{\mathrm{cor}}(Y_i \mid X_i = \overline{X})$, and the latent variables, $\widehat{\mathrm{cor}}(W_i \mid X_i)$, are, respectively,

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

where the variable ordering is square-root estradiol level, log-total gonadotropin level, number of egg cells, and number of embyros. The estimate of $\text{cor}(Y_i \mid X_i)$ is here evaluated at $\overline{X} = \sum_{i=1}^{n} X_i/n$. The estimates indicate substantial positive correlation between the number of egg cells and number 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.[43] 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).

Finally, to illustrate how uncertainty can be quantitified using the approximate likelihood, we construct an approximate 95% confidence interval for Σ_{43} by inverting the proposed approximate likelihood ratio test-statistic numerically. This gives the confidence interval $(0.17, 0.24)$, which corresponds to correlations between 0.62 and 0.8. Confidence intervals for the other parameters could be constructed similarly.

6.2 | Somatic mutations and gene expression in breast cancer

We now focus on jointly modeling common somatic mutations and gene expression measured on patients with breast cancer collected by The cancer genome atlas project (TCGA). A somatic mutation is an alteration in the DNA of a somatic

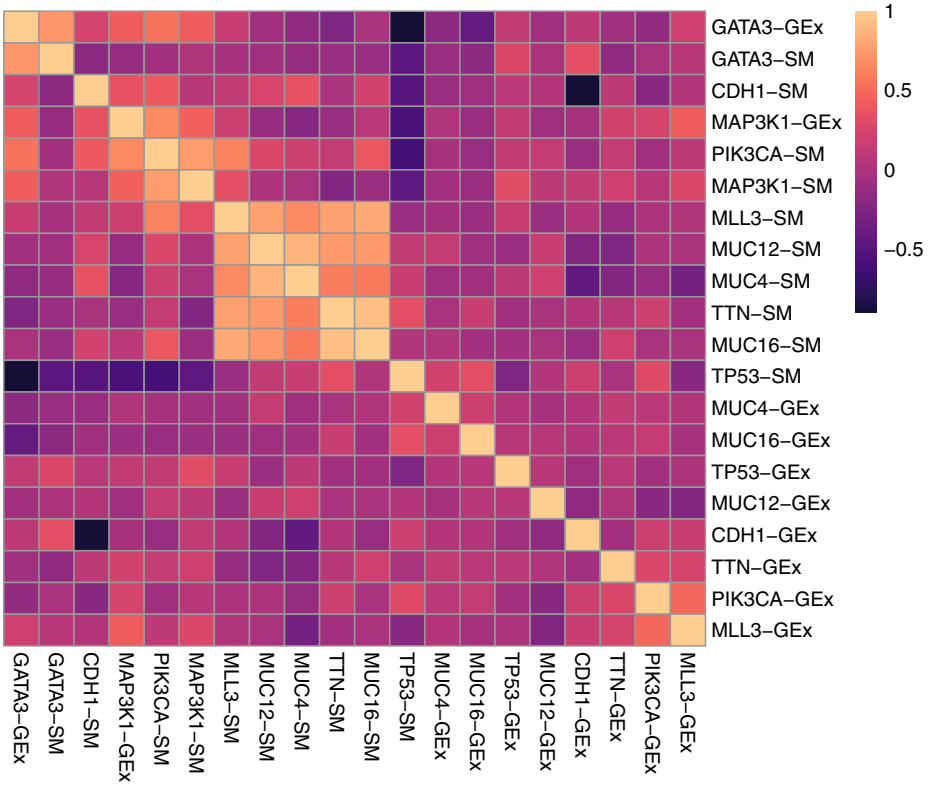

FIGURE 6 Heatmap of the estimated correlation matrix for the $W_i \mid X_i$ in Section 6.2. Suffix -SM for somatic mutations; -GEx for gene expression

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 mutation variables are binary, indicating presence or absence of a somatic mutation in the region of a particular gene. We focus on the ten genes where a somatic mutation was 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. Each patient's age is included as a predictor.

We test the covariance matrix for block-diagonality. Under the null hypothesis, entries of Σ corresponding the correlations between somatic mutations and gene expression measurements are zero (ie, is no correlation between somatic mutations and gene expression). The observed statistic is $T_n = 739$ with 100 degrees of freedom for a P-value $< 10^{-16}$. Figure 6 displays the estimated correlation matrix for the $W_i|X_i$. We observe 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 (eg, TTN, MLL3, MUC4, MUC12, MUC16). However latent variables corresponding to some somatic mutations, for example, those in the region of TP53, exhibit small or even negative correlations with many others (eg, GATA3, CDH1, PIK3CA). Confidence intervals could be constructed as in Example 6.1.

7 | DISCUSSION

We have proposed a likelihood-based method for mixed-type multivariate response regressions. Our method gives an approximate maximum likelihood estimate of an unstructured covariance matrix characterizing the dependence between responses. This is particularly useful in settings where a dependence structure is not suggested by subject-specific knowledge or one wants to discover such a structure using data. To address the computational challenges with estimating an unstructured covariance matrix, we have proposed a new algorithm. The algorithm handles identifiability constraints and it scales well in the dimension of the response vectors.

An advantage of likelihood-based methods compared to ones based on marginal moments, such as for example generalized estimating equations and its extensions,[27-29,44] is the plethora of existing procedures for inference and model selection using likelihoods. For example, Wald tests and standard errors for the coefficient estimates, based on the observed Fisher information, are readily available once maximum likelihood estimates have been computed, as are information criteria. We also used the likelihood to propose a testing procedure for the covariance matrix. On the other hand, it is often computationally expensive to evaluate the likelihood in latent variable models and it can lead to complicated optimizations. We addressed that using a PQL-like approximation and a new algorithm, and showed the resulting approximate maximum likelihood estimates are often useful. Extending our method to other likelihood-approximations is a possibility. For example, the PQL-approximation could be replaced by a Laplace or Monte Carlo approximation. Another potential line of future research is the asymptotic properties of PQL-type estimators, which despite the estimators' apparent practical usefulness, are not fully understood.

Finally, we note some types of mixed-type longitudinal data are handled by our method as-is, while other types would require modifications or would be better analyzed using methods developed for that purpose. For example, our method can be applied when the elements of $Y_i \in \mathbb{R}^r$, $i = 1, \dots, n$, are repeated measures of mixed-type responses on independently sampled patients indexed by i. Modifications may be necessary if one wants to impose a particular structure on Σ, if $n < r$, or if there is dependence between patients.

ACKNOWLEDGEMENTS

The authors thank three reviewers for suggestions which helped improve the manuscript. The authors are grateful to Galin Jones and Adam Rothman for their comments on an earlier version and thank Charles McCulloch for sharing the OAI dataset analyzed in the Supplementary Material. Karl Oskar Ekvall gratefully acknowledges support from FWF (Austrian Science Fund) [P30690-N35]. Aaron J. Molstad's research was supported in part by a grant from the National Science Foundation [DMS-2113589].

DATA AVAILABILITY STATEMENT

Data sharing is not applicable to this article as no new data were created or analyzed in this study.

ORCID

Karl Oskar Ekvall https://orcid.org/0000-0001-8085-4353
Aaron J. Molstad https://orcid.org/0000-0003-0645-5105

REFERENCES

1. Olkin I, Tate RF. Multivariate correlation models with mixed discrete and continuous variables. *Ann Math Stat*. 1961;32(2):448-465. doi:10.1214/aoms/1177705052
2. Poon WY, Lee SY. Maximum likelihood estimation of multivariate polyserial and polychoric correlation coefficients. *Psychometrika*. 1987;52(3):409-430. doi:10.1007/BF02294364
3. Catalano PJ, Ryan LM. Bivariate latent variable models for clustered discrete and continuous outcomes. *J Am Stat Assoc*. 1992;87(419):651-658. doi:10.2307/2290200
4. Cox DR, Wermuth N. Response models for mixed binary and quantitative variables. *Biometrika*. 1992;79(3):441-461. doi:10.1093/biomet/79.3.441
5. Fitzmaurice GM, Laird NM. Regression models for a bivariate discrete and continuous outcome with clustering. *J Am Stat Assoc*. 1995;90(431):845-852. doi:10.2307/2291318
6. Catalano PJ. Bivariate modelling of clustered continuous and ordered categorical outcomes. *Stat Med*. 1997;16(8):883-900. doi:10.1002/(sici)1097-0258(19970430)16:8<883::aid-sim542>3.0.co;2-e
7. Fitzmaurice GM, Laird NM. Regression models for mixed discrete and continuous responses with potentially missing values. *Biometrics*. 1997;53(1):110-122.
8. Gueorguieva RV, Agresti A. A correlated probit model for joint modeling of clustered binary and continuous responses. *J Am Stat Assoc*. 2001;96(455):1102-1112. doi:10.1198/016214501753208762
9. Gueorguieva RV, Sanacora G. Joint analysis of repeatedly observed continuous and ordinal measures of disease severity. *Stat Med*. 2006;25(8):1307-1322. doi:10.1002/sim.2270
10. Yang Y, Kang J, Mao K, Zhang J. Regression models for mixed Poisson and continuous longitudinal data. *Stat Med*. 2007;26(20):3782-3800. doi:10.1002/sim.2776
11. Faes C, Aerts M, Molenberghs G, Geys H, Teuns G, Bijnens L. A high-dimensional joint model for longitudinal outcomes of different nature. *Stat Med*. 2008;27(22):4408-4427. doi:10.1002/sim.3314
12. Rabe-Hesketh S, Skrondal A, Pickles A. Generalized multilevel structural equation modeling. *Psychometrika*. 2004;69(2):167-190. doi:10.1007/bf02295939
13. Sammel MD, Ryan LM, Legler JM. Latent variable models for mixed discrete and continuous outcomes. *J Royal Stat Soc Ser B (Methodol)*. 1997;59(3):667-678.
14. Gueorguieva RV. A multivariate generalized linear mixed model for joint modelling of clustered outcomes in the exponential family. *Stat Model*. 2001;1(3):177-193.
15. de Leon AR, Carrièger KC. General mixed-data model: Extension of general location and grouped continuous models. *Can J Stat*. 2007;35(4):533-548. doi:10.1002/cjs.5550350405
16. Ekvall KO, Jones GL. Consistent maximum likelihood estimation using subsets with applications to multivariate mixed models. *Ann Stat*. 2020;48(2):932-952. doi:10.1214/19-AOS1830
17. Bai H, Zhong Y, Gao X, Xu W. Multivariate mixed response model with pairwise composite-likelihood method. *Stat*. 2020;3(3):203-220. doi:10.3390/stats3030016
18. Kang X, Chen X, Jin R, Wu H, Deng X. Multivariate regression of mixed responses for evaluation of visualization designs. *IISE Trans*. 2021;53(3):313-325. doi:10.1080/24725854.2020.1755068
19. Schall R. Estimation in generalized linear models with random effects. *Biometrika*. 1991;78(4):719-727. doi:10.1093/biomet/78.4.719
20. Breslow NE, Clayton DG. Approximate inference in generalized linear mixed models. *J Am Stat Assoc*. 1993;88(421):9-25.
21. Rabe-Hesketh S, Skrondal A, Pickles A. Reliable estimation of generalized linear mixed models using adaptive quadrature. *Stata J*. 2002;2(1):1-21. doi:10.1177/1536867X0200200101
22. Breslow N. *Whither PQL?* . New York, NY: Springer; 2004:1-22.
23. Knudson C, Benson S, Geyer C, Jones G. Likelihood-based inference for generalized linear mixed models: inference with the R package GLMM. *Stat*. 2021;10(1):e339. doi:10.1002/sta4.339
24. Rabe-Hesketh S, Skrondal A, Pickles A. GLLAMM manual. U. C. Berkeley Division of Biostatistics Working Paper Series; 2004.
25. Bates D, Mächler M, Bolker B, Walker S. Fitting linear mixed-effects models using lme4. *J Stat Softw*. 2015;67(1):1-48. doi:10.18637/jss.v067.i01
26. Jiang J. *Linear and Generalized Linear Mixed Models and Their Applications*. Springer Series in Statistics. New York, NY: Springer-Verlag; 2007.
27. Liang KY, Zeger SL. Longitudinal data analysis using generalized linear models. *Biometrika*. 1986;73(1):13-22. doi:10.1093/biomet/73.1.13
28. Rochon J. Analyzing bivariate repeated measures for discrete and continuous outcome variables. *Biometrics*. 1996;52(2):740-750.
29. Bonat WH, Jørgensen B. Multivariate covariance generalized linear models. *J Royal Stat Soc Ser C (Appl Stat)*. 2016;65(5):649-675. doi:10.1111/rssc.12145
30. McCullagh P, Nelder JA. *Generalized Linear Models*. Boca Raton, FL: Chapman & Hall/CRC Press; 1989.

31. Zellner A. An efficient method of estimating seemingly unrelated regressions and tests for aggregation bias. *J Am Stat Assoc*. 1962;57(298):348-368. doi:10.1080/01621459.1962.10480664
32. Dunson DB. Bayesian latent variable models for clustered mixed outcomes. *J Royal Stat Soc Ser B (Stat Methodol)*. 2000;62(2):355-366.
33. Pinheiro JC, Bates DM. Unconstrained parametrizations for variance-covariance matrices. *Stat Comput*. 1996;6(3):289-296.
34. Ochs P, Chen Y, Brox T, Pock T. iPiano: inertial proximal algorithm for nonconvex optimization. *SIAM J Imaging Sci*. 2014;7(2):1388-1419. doi:10.1137/130942954
35. Beck A, Teboulle M. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM J Imaging Sci*. 2009;2(1):183-202. doi:10.1137/080716542
36. Megginson RE. *An Introduction to Banach Space Theory*. New York, NY: Springer; 1998.
37. Boyle JP, Dykstra RL. A method for finding projections onto the intersection of convex sets in Hilbert spaces. In: Brillinger D, Fienberg S, Gani J, et al., eds. *Advances in Order Restricted Statistical Inference*. Vol 37. New York, NY: Springer; 1986:28-47.
38. Nocedal J, Wright S. *Numerical Optimization*. New York, NY: Springer-Verlag GmbH; 2006.
39. Self SG, Liang KY. Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *J Am Stat Assoc*. 1987;82(398):605-610.
40. Geyer CJ. On the asymptotics of constrained M-estimation. *Ann Stat*. 1994;22(4):1993-2010. doi:10.1214/aos/1176325768
41. Baey C, Cournède PH, Kuhn E. Asymptotic distribution of likelihood ratio test statistics for variance components in nonlinear mixed effects models. *Comput Stat Data Anal*. 2019;135:107-122. doi:10.1016/j.csda.2019.01.014
42. Cannon AR, Cobb GW, Hartlaub BA, et al. *Stat2: Building Models for a World of Data*. New York, NY: W. H. Freeman and Company; 2013.
43. La Marca A, Sunkara SK. Individualization of controlled ovarian stimulation in IVF using ovarian reserve markers: from theory to practice. *Hum Reprod Update*. 2013;20(1):124-140.
44. Ziegler A. *Generalized Estimating Equations. No. 204 in Lecture notes in Statistics*. New York, NY: Springer; 2011.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

How to cite this article: Ekvall KO, Molstad AJ. Mixed-type multivariate response regression with covariance estimation. *Statistics in Medicine*. 2022;41(15):2768-2785. doi: 10.1002/sim.9383