Bayesian Variable Selection for Gaussian Copula Regression Models

Angelos Alexopoulos and Leonardo Bottolo

Department of Statistical Science, University College London, London, UK; Department of Medical Genetics, University of Cambridge, UK; The Alan Turing Institute, London, UK; MRC Biostatistics Unit, University of Cambridge, Cambridge, UK

ABSTRACT

We develop a novel Bayesian method to select important predictors in regression models with multiple responses of diverse types. A sparse Gaussian copula regression model is used to account for the multivariate dependencies between any combination of discrete and/or continuous responses and their association with a set of predictors. We use the parameter expansion for data augmentation strategy to construct a Markov chain Monte Carlo algorithm for the estimation of the parameters and the latent variables of the model. Based on a centered parameterization of the Gaussian latent variables, we design a fixed-dimensional proposal distribution to update jointly the latent binary vectors of important predictors and the corresponding nonzero regression coefficients. For Gaussian responses and for outcomes that can be modeled as a dependent version of a Gaussian response, this proposal leads to a Metropolis-Hastings step that allows an efficient exploration of the predictors’ model space. The proposed strategy is tested on simulated data and applied to real datasets in which the responses consist of low-intensity counts, binary, ordinal and continuous variables.

1. Introduction

The identification of important predictors in linear and non-linear regression models is one of the most frequently studied questions in statistical theory. In Bayesian statistics, this problem is known as Bayesian variable selection (BVS) and for Gaussian responses, there is an extensive literature for an efficient detection of important predictors in both single- and multi-response regression models. In the case of a single response, a list of relevant papers includes, but is not limited to, George and McCulloch (1997), Liang et al. (2008), Guan and Stephens (2011) and Ročková and George (2018), while Brown, Vannucci, and Fearn (1998), Holmes, Denison, and Mallick (2002) deal with the same problem in the multi-response case. BVS in single-response nonlinear regression models has also received great attention. In Dellaportas, Forster, and Ntzoufras (2002) and Forster, Gill, and Overstall (2012) the corresponding methods are reviewed and advances are proposed.

More recently, there has been an increasing interest in the joint analysis of outcomes of diverse types, for instance, continuous, binary, categorical and count data, given their availability from studies involving multivariate data, see, for example, Hoff (2007), Murray et al. (2013), Zhang et al. (2015), and Bhadra, Rao, and Baladandayuthapani (2018). In regression analysis, the most popular model used to account for the multivariate dependencies between any combination of discrete and/or continuous responses is the Gaussian copula regression (GCR) model (Song, Li, and Yuan 2009) in which each response is associated with a (potentially different) set of predictors. When only Gaussian responses are considered, this is known as Seemingly Unrelated Regression (SUR) model (Zellner 1962). Recent contributions for sparse Bayesian SUR models include, for instance, Wang (2010) and Deshpande, Ročková, and George (2019). Bayesian methods for the estimation of the regression coefficients of the GCR model with a fixed set of predictors have also been proposed (Pitt, Chan, and Kohn 2006). Despite the growing Bayesian literature regarding an efficient selection of important predictors, to the best of our knowledge, variable selection for the GCR model has not been attempted. We propose here the first fully Bayesian approach for model selection in regression models with multiple responses of diverse types.

The main obstacle of the application of BVS in single-response non-linear models, as well as in the SUR model and the GCR model with responses of diverse types, is the non-tractability of the marginal likelihood. To overcome this problem, Markov chain Monte Carlo (MCMC) algorithms are based either on Laplace approximation, see for example Bové and Held (2011), or on a Metropolis-Hastings (M-H) step in which the dimension of the proposal distribution is not fixed at each iteration (Forster, Gill, and Overstall 2012). The latter is an application of the reversible jump algorithm (Green 1995) which is known to experience a low acceptance rate when the transdimension proposal distribution is not devised carefully, resulting in MCMC samplers with poor mixing (Brooks, Giudici, and Roberts 2003). In addition, in current applications, the number of predictors is often very large and any MCMC algorithm for BVS in both linear and nonlinear regression has to be designed carefully in order to explore successfully the ultra-high dimensional...
model space consisting of all the possible subsets of predictors (Lamnisos, Griffin, and Steel 2009; Bottolo and Richardson 2010).

The main contribution of this article is the development of a Bayesian approach for the joint update of the latent binary vector of important predictors and the corresponding vector of non-zero regression coefficients for each response of the GCR model. By using the proposed strategy, we perform BVS without any approximation. We also avoid the reversible jump algorithm by using the Gaussian latent variables of the GCR model to construct a proposal distribution defined on a fixed-dimensional space. For Gaussian responses and for the outcomes that can be modeled as a dependent version of a Gaussian response, for example, the Probit model for binary data, the designed approximations can be used. We also avoid the reversible jump algorithm by using the proposed strategy, we perform BVS without any approximation. We also avoid the reversible jump algorithm by using the Gaussian latent variables of the GCR model to construct a proposal distribution defined on a fixed-dimensional space. For Gaussian responses and for the outcomes that can be modeled as a dependent version of a Gaussian response, for example, the Probit model for binary data, the designed approximations can be used.

Modeling the dependence amongst the responses is another key aspect in GCR. Until recently devising an efficient MCMC algorithm for a structured (constrained) covariance matrix that includes the identifiability conditions for the non-linear responses has been a difficult task. Here, we follow the solution proposed by Talhouk, Doucet, and Murphy (2012) and specify a conjugate prior on the correlation matrix. We use the parameter expansion for data augmentation (Liu and Wu 1999; Van Dyk and Meng 2001) to expand the correlation into a covariance matrix. We also adopt the idea of covariance selection to obtain a parsimonious representation of the dependence amongst the responses. Based on the theory of decomposable Gaussian graphical models (Lauritzen 1996), we use the hyper-inverse Wishart distribution as the prior density for the covariance matrix. This prior allows some of the off-diagonal elements of the inverse covariance matrix to be identical to zero and to estimate the conditional dependence pattern of the observations (Webb and Forster 2008).

We tested the performance of our model, Bayesian Variable Selection for Gaussian Copula Regression (BVSGCR), in a comprehensive simulation study and compared the performance of our new approach with conventional Bayesian methods for the selection of important predictors in single-response (linear and non-linear) regression models, see Holmes and Held (2006), Frühwirth-Schnatter et al. (2009) and Dvorzak and Wagner (2016).

We also applied the proposed method to two real datasets. The first dataset includes a combination of nine continuous, binary and ordered categorical responses. These responses are phenotypic traits of a rare disorder called Ataxia-Telangiectasia. We analyzed one of the largest cohorts of patients, consisting of 46 individuals affected by the disease (Schon et al. 2019). Our model borrows information across the outcomes in order to identify important associations between the responses and a set of genetic and immunological predictors that have been collected in the same study. The second dataset consists of four counts and one ordered categorical response which are measured in 122 individuals suffering from Temporal Lobe Epilepsy (Johnson et al. 2016). Our interest lies in the identification of associations between the responses and 162 correlated genes that have been identified in a recent gene-network analysis related to cognition abilities and epilepsy (Johnson et al. 2016). Finally, in both real datasets, we compared the predictive ability of the BVSGCR model with widely-used single-response linear and non-linear sparse regression models.

The rest of the article is organized as follows. In Section 2 we provide a brief presentation of the GCR model and the prior distributions on the regression coefficients and the correlation structure. In Section 3 we describe the novel MCMC algorithm that we propose for BVS when a combination of discrete and/or continuous responses is considered. Section 4 presents the results of the simulation study and in Section 5 we apply the proposed model on two real datasets with missing values in the outcome variables which led to a straightforward modification of the designed MCMC algorithm. Finally, in Section 6 we conclude with a short discussion.

2. Gaussian Copula Regression Model

In the following, all vectors, in bold font, are understood as column vectors and the superscript “T” is used to denote the transpose of a vector or a matrix. Matrices are also indicated in bold font. The lower-case notation will be reserved for the observations with the corresponding random variables in capital letters.

2.1. Gaussian Copulas

An m-variate function \( C(u_1, \ldots, u_m) \), where \( C : [0, 1]^m \rightarrow [0, 1] \), is called a copula if it is a continuous distribution function and each marginal is a uniform distribution function on \([0, 1]\). Sklar (1959) proved that any joint cumulative distribution function (cdf) of continuous random variables can be completely specified by its marginal distributions and a unique copula \( C \). If \( F_1(\cdot), \ldots, F_m(\cdot) \) are the marginal cdfs of a combination of \( m \) continuous and discrete random variables \( Y_1, \ldots, Y_m \), their joint cdf can be specified through a specific copula function \( C \) as

\[
F_{Y_1, \ldots, Y_m}(y_1, \ldots, y_m) = C(F_{Y_1}(y_1), \ldots, F_{Y_m}(y_m)).
\]

A copula function which is commonly used for modeling the dependence structure of any combinations of continuous and discrete variables is the Gaussian copula, see for example Hoff (2007) and Murray et al. (2013). The Gaussian copula \( C \) is specified through the function

\[
C(u_1, \ldots, u_m; R) = \Phi_m(\Phi^{-1}(u_1), \ldots, \Phi^{-1}(u_m); R),
\]

where \( \Phi_m(\cdot; R) \) is the cdf of an \( m \)-variate Gaussian distribution with zero mean vector and correlation matrix \( R \) and \( \Phi^{-1}(\cdot) \) is the inverse of the univariate standard normal cdf. Thus, taking in Equation (1) \( u_k = F_{Y_k}(y_k) \), for each \( k = 1, \ldots, m \), we specify the cdf of \( Y = (Y_1, \ldots, Y_m) \) to be the Gaussian copula function. Song (2000) proves that the density of the Gaussian copula is

\[
c(u_1, \ldots, u_m; R) = |R|^{-\frac{1}{2}} \exp \left\{ -\frac{1}{2} (\tilde{z}^T R^{-1} \tilde{z} - \tilde{z}^T \tilde{z}) \right\},
\]

where \( \tilde{z} \) is an \( m \)-dimensional vector and \( \tilde{z}_k = \Phi^{-1}(u_k), k = 1, \ldots, m \), is known as the normal score which follows marginally.
the standard Gaussian distribution. Then, \( \tilde{Z} \) is an \( m \)-variate Gaussian distribution with zero mean vector and correlation matrix \( R \), that is, \( \tilde{Z} \sim N_m(0, R) \). If all the marginal distributions are continuous, the matrix \( R \) can be interpreted as the correlation matrix of the elements of \( Y \) and zeros in its inverse imply the conditional independence among the corresponding elements of \( Y \). However, in the presence of discrete random variables, the notion of conditional independence has to be interpreted with care; zeros in \( R^{-1} \) imply that the observed variables are independent conditionally only on the latent variables (Webb and Forster 2008). Note also that if \( R \) is the identity matrix the elements of \( Y \) can be considered to be independent despite the presence of discrete variables, see Song (2000), Song, Li, and Yuan (2009) and Talhouk, Doucet, and Murphy (2012) for a detailed discussion.

### 2.2. Regression Model

Let \( Y = (y_1^T, \ldots, y_m^T) \) be the \((n \times m)\)-dimensional matrix of observed data where each \( y_i = (y_{i1}, \ldots, y_{im}) \) consists of any combination of \( m \) discrete and/or continuous responses. Moreover, let \( x_{jk} \) be the \( pk \)-dimensional vector of predictors for the \( i \)th sample and the \( k \)th response. To model the distributional form of each \( y_{ik} \), we specify its cdf \( F_k(y_{ik}; x_{jk}, \beta_k, \theta_k) \) to be the cdf of any parametric distribution. Our notation emphasizes the dependence of each \( y_{ik} \) on the \( pk \)-dimensional response-specific vector of predictors \( x_{jk} \), the associated regression coefficients \( \beta_k \) and the response-specific parameters \( \theta_k \). The GCR model is described by the transformation

\[
y_{ik} = h_{ik}^{-1}(\tilde{Z}_{ik}), \quad \tilde{Z}_i \sim N_m(0, R),
\]

where \( \tilde{Z}_{i1}, \ldots, \tilde{Z}_{im} \) are realizations from \( \tilde{Z}_i \) and \( h_{ik}(\cdot) = \Phi^{-1}[F_k(y_{ik}; x_{jk}, \beta_k, \theta_k)] \) for each \( i = 1, \ldots, n \) and \( k = 1, \ldots, m \).

By assuming that each \( F_k(\cdot) \) is a member of the exponential family, we obtain the multivariate Generalized Linear Model presented in Song, Li, and Yuan (2009) as the multivariate extension of the well-known single-response generalized linear model (McCullagh and Nelder 1989) with \( \theta_k \) the specific vector of parameters for the \( k \)th response. The SUR model is a special case of Equation (3) when all margins are univariate Gaussian with mean \( x_{jk}^T \beta_k \) and variance \( \theta_k \) and \( R \) is the correlation matrix. The multi-response Probit regression model of Chib and Greenberg (1998) is obtained from Equation (3) by specifying each margin to be the cdf of a Bernoulli random variable with probability of success \( \Phi(x_{jk}^T \beta_k) \). Finally, setting \( F_k(y_{ik}; x_{jk}, \beta_k, \theta_k) = \Phi(\theta_{kc} - x_{jk}^T \beta_k), c = 1, \ldots, C_k - 1 \), with \( C_k \) the number of categories, Equation (3) becomes a regression model for a combination of binary and ordinal observations and \( \theta_k = (\theta_{k1}, \ldots, \theta_{kC_k-1}) \) consists of the cut-points for the \( k \)th ordinal observation (McCullagh 1980). In both multi-response Probit and ordinal regression models the matrix \( R \) is in the correlation form for identifiability conditions (Chib and Greenberg 1998).

Equation (3) implies that the joint likelihood function of the observations \( Y \) conditional on the correlation matrix \( R \), the regression coefficients \( B = \text{vec}(\beta_1, \ldots, \beta_m) \) and the parameters \( \Theta = \text{vec}(\theta_1, \ldots, \theta_m) \) is an intractable function of noncomputable high-dimensional integrals (Song, Li, and Yuan 2009).

### 2.3. Prior Distributions

In this section, we specify the prior distributions on the regression coefficients \( B \), the associated sparsity prior on the inclusion probability and the correlation matrix \( R \) of the model. By noting that for different choices of the marginal cdfs \( F_k(\cdot), k = 1, \ldots, m \), we have different vectors of parameters \( \theta_k \), we will assign their prior distributions differently in each simulated and real data example. See Supplementary Section S.1 for details.

#### 2.3.1. Variable Selection

We use a hierarchical nonconjugate model to assign a prior distribution on the regression coefficients of the GCR model defined in Equation (3). A point mass at zero is specified on the regression coefficients of the unimportant predictors, whereas a Gaussian distribution is assigned to the nonzero regression coefficients (George and McCulloch 1997). By using the binary latent vector \( y_k = (y_{k1}, \ldots, y_{kp_k}), k = 1, \ldots, m \), where \( y_{kj} \) is 1 if, for the \( j \)th predictor and the \( k \)th response, the regression coefficient is different from zero and 0 otherwise, we assume that for each \( k \)

\[
\beta_{kj} | y_{kj} \overset{\text{iid}}{\sim} (1 - y_{kj}) \delta_0 + y_{kj} N(0, v), \quad j = 1, \ldots, p_k,
\]

\[
y_{kj} | \pi_k \overset{\text{iid}}{\sim} \text{Ber}(\pi_k), \quad j = 1, \ldots, p_k,
\]

\[
\pi_k \sim \text{Beta}(a_k, b_k),
\]

where \( \delta_0 \) denotes a point mass at zero and \( v \) is a fixed value. It is common practice to standardize the predictor variables, taking \( v = 1 \) in order to place appropriate prior mass on reasonable values of the nonzero regression coefficients (Hans, Dobra, and West 2007). Integrating out \( \pi_k \), it is readily shown that marginally

\[
p(y_k) = \frac{B(a_k + |y_k|, b_k + p_k - |y_k|)}{B(a_k, b_k)},
\]

where \( |y_k| = \sum_{j=1}^{p_k} y_{kj} \). The hyperparameters \( a_k \) and \( b_k \) can be chosen using prior information about the number of important covariates associated with the \( k \)th response and its variance. See (Kohn, Smith, and Chan 2001) for further details on the elicitation of the hyper-parameters \( a_k \) and \( b_k \) in Equation (5).

The sparsity prior (5) is a common choice in single-response BVS as well as in the sparse SUR model (Wang 2010). When the predictors are common to all responses, the primary inferential question may shift to the identification of key predictors that exert their influence to a large fraction of responses at the same time. When a large number of responses are regressed independently on the same set of predictors, Richardson, Bottolo, and Rosenthal (2010) proposed to decompose the \( a \) priori “cell” inclusion probability into its marginal effects, that is, \( \pi_{kj} = \pi_k \times \rho_j, \pi_{kj} \in [0, 1] \). The idea behind this decomposition is to control the level of sparsity for each response \( k \) through a suitable choice of the hyperparameters \( a_k \) and \( b_k \) as in Equation (5), while \( \rho_j \sim \text{Ga} (c, d) \) captures the relative “propensity” of predictor \( j \) to influence several responses at the same time. For a detailed discussion regarding the effect of the prior \( p(\rho) \) on BVS in multiple-response regression models, see Ruffieux et al. (2020).
Let $g_{\ell}$ priordistribution for the covariance matrix $\Sigma$ ensure the positiveness of $X$. Let $\pi$ of the adjacency matrix $G$.

### 2.3.2. Correlation Matrix $R$ and Adjacency Matrix $G$

To specify a prior distribution on the correlation matrix $R$, we follow Talhouk, Doucet, and Murphy (2012) and use the parameter expansion for data augmentation strategy (Liu and Wu 1999) to expand the correlation matrix into a covariance matrix. This choice allows us to specify conjugate prior distributions on the resulting covariance matrix. In particular, we use the hyper-inverse Wishart (Dawid and Lauritzen 1993) to allow for zero entries in the inverse of the covariance matrix. Details can be summarized as follows.

First, to expand $R$ into a covariance matrix, we define the transformation $\tilde{W} = \tilde{Z}D$, where $\tilde{Z}$ is the $n \times m$ matrix of the Gaussian latent variables and $D$ is an $m \times m$ diagonal matrix with elements $\delta_{k}k = 1, \ldots, m$. Then,

\[
\text{vec}(\tilde{W}) \sim N_{nm}(0, \Sigma \otimes I_{n}),
\]

where $\Sigma = D\Sigma D$ and $I_{n}$ is a diagonal matrix of dimension $n$ that encodes the independence assumption amongst the observations. Then, a conjugate prior distribution can be assigned on $\Sigma$ and updated at each iteration of the MCMC algorithm before it is projected back to $R$ using the inverse transformation $R = D^{-1}\Sigma D^{-1}$. We also use the theory of decomposable models to perform a conjugate analysis of the covariance structure of the model since the hyper-inverse Wishart distribution is a conjugate prior distribution for the covariance matrix $\Sigma$ with respect to the adjacency matrix $G$ of a decomposable graph $\mathcal{G}$. The diagonal elements in the adjacency matrix are always restricted to be 1 to ensure the positive definiteness of $G$.

Second, we assign the following prior structure on $G$ and $D$. Let $g_{\ell}, \ell = 1, \ldots, m(m - 1)/2$, be the binary indicator for the presence of the $\ell$th-off diagonal edge in the lower triangular part of the adjacency matrix $G$ of the decomposable graph $\mathcal{G}$. We assume that

\[
g_{\ell} \sim \text{Ber}(\pi_{G}), \quad \pi_{G} \sim \text{Unif}(0, 1), \quad \ell = 1, \ldots, m(m - 1)/2, \quad G \in \mathcal{G}.
\]

For a detailed discussion regarding compatible priors on $G$ for decomposable graphs, see Bornn and Caron (2011).

We denote by $p(G)$ the induced marginal prior on the adjacency matrix and define the joint distribution of $D, R$ and $G$ as

\[
p(D, R, G) = p(D|R)p(R|G)p(G),
\]

where

\[
\delta_{k}^{2} \sim \text{IGam}((m + 1)/2, k^{2}/2), \quad k = 1, \ldots, m
\]

with $x^{kk}$ the $k$th diagonal element of $R^{-1}$. Assuming a uniform prior for $R|G$, it can be shown that, for a decomposable graph $\mathcal{G}$, $\Sigma|G \sim \text{HIW}_{G}(2, I_{m})$ (Talhouk, Doucet, and Murphy 2012).

### 3. MCMC Sampling Strategy

We are interested in sampling from the joint posterior distribution $p(B, \Gamma, \Theta, \tilde{Z}, D, R, G|Y)$. To draw samples from the specified model, we design a novel MCMC algorithm which proceeds as follows. We first update the regression coefficients $B$, the latent binary matrix $\Gamma$, the parameters $\Theta$ and the Gaussian latent variables $\tilde{Z} = (\tilde{Z}_{1k}, \ldots, \tilde{Z}_{mk})$. In the last step, we draw the correlation matrix $R$ and the adjacency matrix $G$ from their full conditional distributions.

**Algorithm 1** MCMC algorithm for sampling from the joint distribution $p(B, \Gamma, \Theta, \tilde{Z}, D, R, G|Y)$

1: Set the number of iterations $S$
2: for $s = 1, \ldots, S$ do
3: for $k = 1, \ldots, m$ do
4: Sample from $p(\beta_{k}, y_{k}|\beta_{-k}, \tilde{Z}_{-k}, D, R)$
5: Sample from $p(\theta_{k}|y_{k}, \beta_{k}, y_{k}, \tilde{Z}_{-k}, D, R)$
6: Sample from $p(\tilde{Z}_{k}|y_{k}, \beta_{k}, y_{k}, \theta_{k}, \tilde{Z}_{-k}, D, R)$
7: end for
8: Sample from $p(D, R|Y, \tilde{Z}, G)$
9: Sample from $p(G|Y, \tilde{Z}, D, R)$
10: end for

For the large majority of responses, drawing $(\beta_{k}, y_{k})$ in Step 4 is complicated due to the unavailability of the full conditional distributions in closed-form expression and a M-H step is therefore required. However, because of the uncertainty associated with the latent binary vector $y_{k}$, the proposal distribution requires a different dimension at each MCMC iteration. In this framework, a commonly used tool is the Reversible Jump algorithm (Green 1995), although it may experience a low acceptance rate, resulting in a MCMC sampler with poor mixing (Brooks, Giudici, and Roberts 2003; Lamnisos, Griffin, and Steel 2009).

To avoid the reversible jump algorithm, we design a proposal distribution defined on a fixed-dimensional space that can be used for the joint update of $\beta_{k}$ and $y_{k}$. It is worth noticing that we conduct the update of $(\beta_{k}, y_{k})$ by first integrating out the latent variables $\tilde{Z}_{k}$. This accelerates the convergence of the proposed MCMC algorithm since conditioning on $\tilde{Z}_{k}$ induces many restrictions on the admissible values of $(\beta_{k}, y_{k})$, see Pitt, Chan, and Kohn (2006). Drawing samples from the posterior distributions of the remaining parameters and the Gaussian latent variables of the model can be conducted by using standard MCMC algorithms which we also describe briefly in this section.

#### 3.1. Proposal Distribution for Variable Selection

To sample jointly the regression coefficients $\beta_{k}$ and the latent binary vector $y_{k}$ for each $k = 1, \ldots, m$ (step 4 Algorithm 1), we...
design a M-H step that targets the distribution with density

\[ p(\beta_k, y_k | \theta_k, Z_{-k}, D, R) \propto p(y_k | \beta_k, y_k, \theta_k, \tilde{Z}_{-k}, D, R) p(\beta_k | y_k) p(y_k) \]  

where \( p(\beta_k | y_k) \) and \( p(y_k) \) are the prior densities on \( \beta_k \) and \( y_k \) as defined in Equations (4) and (6), respectively.

In the case of a continuous response \( y_k \), we have from Equation (2) that

\[ p(y_k | \beta_k, y_k, \theta_k, \tilde{Z}_{-k}, D, R) \propto \exp \left\{ \frac{1}{2} (1 - r_{kk}) \sum_{i=1}^{n} z_{ik}^{2} \right\} - \sum_{i=1}^{n} \sum_{l=1, l \neq k}^{m} r_{kl} \tilde{z}_{ik} \tilde{z}_{il} + \sum_{i=1}^{n} \log f_{k}(y_{ik}; x_{ik}, \beta_y, \theta_k) \right\}, \tag{11} \]

where we condition on \( Z_{-k} \) with an abuse of notation (Pitt, Chan, and Kohn 2006) since \( \tilde{z}_{ik} = \Phi^{-1}(F_k(y_{ik}; x_{ik}, \beta_y, \theta_k)) \) is a shorter expression of the transformation in Equation (3), \( r_{kk} = (R^{-1})_{kk} \) and \( f_k(y_{ik}; x_{ik}, \beta_y, \theta_k) \) denotes the probability density function of \( F_k(y_{ik}; x_{ik}, \beta_y, \theta_k) \).

If \( y_k \) is a discrete response, we have that

\[ p(y_k | \beta_k, y_k, \theta_k, \tilde{Z}_{-k}, D, R) = \prod_{i=1}^{n} \left( \frac{\mu_{ik} - \tilde{\mu}_{i,k-1}}{\tilde{\sigma}_{i,k-1}} \right) - \Phi \left( \frac{\ell_{ik} - \tilde{\mu}_{i,k-1}}{\tilde{\sigma}_{i,k-1}} \right), \tag{12} \]

where \( \ell_{ik} = \Phi^{-1}(F_k(y_{ik} - 1; x_{ik}, \beta_y, \theta_k)) \) and \( u_{ik} = \Phi^{-1}(F_k(y_{ik}; x_{ik}, \beta_y, \theta_k)) \) with \( \tilde{\mu}_{i,k-1} = R_{ik-1} \tilde{z}_{ik-1} \) and \( \tilde{\sigma}_{i,k-1}^{2} = 1 - R_{ik-1} R_{k-1}^{T} \). Thus, the conditional density in Equation (10) will be intractable for the majority of the continuous and for all the discrete distributions that can be used for the marginal modeling of the 4th response.

Instead of relying on the reversible jump algorithm or the Laplace approximation to sample from Equation (10), we utilize a M-H step with a fixed-dimensional proposal distribution obtained by reparameterizing the Gaussian latent variables \( \tilde{Z}_k \). More precisely, for each \( i = 1, \ldots, n \) and \( k = 1, \ldots, m \), we set

\[ Z_{ik} = x_{ik}^{T} \beta_y + \delta_{ik} \tilde{z}_{ik} \quad \text{if} \quad y_k \text{ is continuous,} \]
\[ Z_{ik} = x_{ik}^{T} \beta_y + \tilde{z}_{ik} \quad \text{if} \quad y_k \text{ is discrete,} \tag{13} \]

where \( Z_{ik} \) is marginally distributed with mean \( \mu_{ik} = x_{ik}^{T} \beta_y \) and variance either \( \sigma^2_{ik} \) (if continuous) or \( \alpha_{ik}^2 = 1 \) (discrete). Thus, the vector of realizations \( z_k \) is a sufficient statistics for \( (\beta_k, y_k) \), whereas \( \tilde{z}_k \) is ancillary (Yu and Meng 2011). Irrespective of the type of the response, Equation (13) creates a link between the Gaussian Copula model, where the Gaussian latent variables are noncentered, and BVS that requires a centered parameterization. With this transformation, we also aim to keep the advantages of using (data augmentation) centered auxiliary variables when performing BVS (Holmes and Held 2006). In particular, the fact that the vector \( z_k \) retains information about the likelihood is key since, regardless of the response’s type, it allows an effective update of \( \beta_k \) given a change in the predictors set.

To construct a proposal distribution for a M-H step that targets Equation (10), we replace the intractable density \( p(y_k | \beta_k, y_k, \theta_k, Z_{-k}, R) \) with the Gaussian density of the latent variable \( Z_k \) conditioned on \( \tilde{Z}_{-k} \). Exploiting the fact that \( \beta_k \) is quadratic in \( p(z_k | \beta_k, y_k, Z_{-k}, D, R) \), given a candidate value of the latent binary vector \( y_k^* \), we propose the nonzero regression coefficients \( \beta_k^* \) from the distribution with density

\[ q(\beta_k^* | y_k^*, z_k, \tilde{Z}_{-k}, D, R) \propto p(z_k | \beta_k, y_k^*, \tilde{Z}_{-k}, D, R) p(\beta_k^* | y_k^*) \propto N(y_k^* | \beta_k^*, \sigma^{-2}) \]  

where \( \sigma^{-2}_{i,k} \) is the diagonal element of the conditional covariance matrix of \( Z_{ik} | \tilde{Z}_{-k} \).

Finally, since the main focus of this article is to provide a proposal distribution for the regression coefficients given the latent binary vector, any proposal distribution for \( y_k^* \) can be used. Here, we use a modified version of the proposal distribution designed by Guan and Stephens (2011). This proposal distribution...
makes efficient use of the inexpensive evaluation of the marginal association between each response and the predictors in order to propose a new latent vector $y^*_k$. See Supplementary Section S.2 for a detailed description.

### 3.2. Sampling the Gaussian Latent Variables

Next, we describe the sampling strategy for the Gaussian latent variables (Step 6 Algorithm 1). If the $k$th response is continuous then the transformation (3) is a one-to-one transformation. In this case $\tilde{z}_k$ is updated deterministically by setting $\tilde{z}_{ik} = \Phi^{-1}(F_k(y|ik;\mu_{ik},\sigma^2_{ik}|k))$ for each $i = 1, \ldots, n$ and $k = 1, \ldots, m$. For a discrete response $y_k$, we have that

$$p(\tilde{z}_{ik}|y_k,\beta_k,\gamma_k,\theta_k,\tilde{Z}_{-k},D,R) \propto N(\tilde{z}_{ik}; \hat{\mu}_{ik|k},\hat{\sigma}^2_{ik|k})$$

(17)

$$I(\tilde{z}_{ik} \in (\ell_{ik},u_{ik}])$$

where $\hat{\mu}_{ik|k},\hat{\sigma}^2_{ik|k},\ell_{ik}$ and $u_{ik}$ are defined in Equation (12). Therefore, each $\tilde{z}_k$ has to be sampled from the Gaussian distribution $N(\mu_{ik|k},\sigma^2_{ik|k})$ truncated on the interval $(\ell_{ik},u_{ik}]$.

### 3.3. Sampling the Matrix $D$, the Correlation Matrix $R$ and the Adjacency Matrix $G$

To update $(D,R)$ and $G$ in the designed MCMC Algorithm 1 (steps 8 and 9), we follow Talhouk, Doucet, and Murphy (2012) and work in the space of the scaled Gaussian latent variables $\tilde{W} = \tilde{Z}D$. In practice, we obtain samples from $p(\Sigma|\tilde{W},G)$ and transform them using the inverse transformation $R = D^{-1}\Sigma D^{-1}$ where $D$ is sampled from its prior distribution in Equation (9). To sample the adjacency matrix $G$ from $p(G|Y,\tilde{Z},D,R)$, we target the distribution with density

$$p(G|\tilde{W}) = \frac{p(\tilde{W}|G)p(G)}{\sum G p(\tilde{W}|G)p(G)},$$

(18)

where the summation in the denominator is over all the decomposable graphs $G$, $p(G)$ is defined through Equation (8) and $p(\tilde{W}|G)$ can be computed analytically due to the tractability of the hyper-inverse Wishart distribution, see Supplementary Section S.5 for further details. We sample from Equation (18) by using a M-H step in which, conditionally on the current adjacency matrix $G$, a new graph is proposed by adding or deleting an edge between two vertices whose index has been chosen randomly between the vertices that belong to a decomposable graph. The proposed graph is then accepted or rejected using the accept/reject mechanism of the M-H step which targets the density in Equation (18). Finally, conditionally on $\tilde{W}$ and the updated adjacency matrix $G$, we sample $\Sigma$ from its conditional distribution $HIW_G(2+n,\mu_m + \tilde{W}^T\tilde{W})$.

**Remark 3.** The choice of a decomposable graphical model can be relaxed to include non-decomposable graphs although at a higher computational cost. Within our framework, posterior samples of the adjacency matrix can be obtained by using the BDgraph algorithm proposed by Mohammadi and Wit (2019) directly on the space of the scaled Gaussian latent variables $\tilde{W}$ where the problem of the intractable normalizing constants appearing in $p(\tilde{W}|G)$ for nondecomposable graphs is circumvented by using the solutions proposed by Wang and Li (2012) and Lenkoski (2013). As a note of caution, by specifying a non-decomposable graphical model, (all else unchanged) the induced prior on $R$ for the incomplete prime components of $G$ may not be uniform. See Supplementary Section S.5 for a detailed discussion.

### 3.4. Sampling the Response-specific Parameters

Sampling the response-specific parameters $\Theta$ depends on the marginal cdfs of the BVSGCR model. If $F_k(\cdot)$ is the cdf of a normal distribution, since $\theta_k = \delta^*_k$, the posterior samples of $\theta_k$ are obtained as a by-product of procedure described above. If the margins are ordinal or negative binomial, as in the real examples considered below, the response-specific parameters are sampled using a M-H step.

### 4. Simulation Study

In this section, we compare the performance of the proposed Bayesian variable selection for Gaussian Copula Regression (BVSGCR) model with widely-used methods for BVS in single-response linear and nonlinear regression models. We tested our multivariate method in two simulated datasets consisting of a combination of Gaussian, binary and ordinal responses, Section 4.2 and Gaussian and count responses, Section 4.3, respectively.

We used the marginal posterior probability of inclusion (MPPPI) (George and McCulloch 1997) to assess the predictor-response association, defined as the frequency a particular predictor is included in a model during the MCMC exploration. To illustrate the performance of the different methods, we used the receiver operating characteristic (ROC) curve. For a given response, the ROC curve plots the proportion of correctly detected important predictors (true positive rate - TPR) against the proportion of misidentified predictors (false positive rate - FPR) over a range of specified thresholds for the MPPPI. To take into account the Monte Carlo error, we reported the mean of TPR and FPR over the simulated replicates for each scenario considered along with the corresponding averaged areas under the curve and their standard deviations. We also assess the accuracy of the nonzero regression coefficients’ posterior credible intervals by using a modified version of the interval score described in Gneiting and Raftery (2007).

### 4.1. Data Generation

To generate the correlated predictors, we followed Rothman, Levina, and Zhu (2010) and simulated, for each $i = 1, \ldots, n$ and $k = 1, \ldots, m$, $x_{ik} \sim N_{p_k}(0,S)$, where $S_{ij} = \theta_{ij}^{0.7|j-i|}$ is the $(j,f)$th element of $S$, $j,f = 1, \ldots, p_k$, implying the same unit marginal variance. We also assumed that, for all responses, we had the same set of available predictors.

We simulated a sparse vector of regression coefficients $B$ and a sparse inverse correlation matrix $R^{-1}$ according to the structure described in Section 2.3. More precisely, we first constructed the $p \times m$ matrix $B = \Gamma_1 \odot \Gamma_2 \odot B_3$, where $p = \sum_1^m p_k$ and $\odot$ denotes the Hadamard matrix product, as follows. Each cell of $\Gamma_1$ has independent Bernoulli entries with
success probability $\pi_1$, $\Gamma_2$ has rows that are either all ones or all zeros and $B_3$ consists of independent draws from $N(b, s^2)$. The decision regarding the zero rows in $\Gamma_2$ has to be made using $p$ independent Bernoulli variables with probability of success $\pi_2$. As noted by Rothman, Levina, and Zhu (2010), using this simulation scheme, $(1 - \pi_2)p$ predictors are expected to be irrelevant for all the responses and each relevant predictor will be associated on average with $\pi_1m$ responses. We set $\beta_k$ to be the $k$th column of $B$. The choice of the parameters $\pi_1, \pi_2, b$ and $s^2$ will be different for each simulated scenario and it is summarized in Table 1. Finally, we used the correlation matrix of the autoregressive model of order one to simulate $\tilde{Z}_{i \sim iid} \sim \mathcal{N}_m(0, R)$ for each $i = 1, \ldots, n$ and we set $R_{kk'} = 0.8^{|k-k'|}$ for the $(k, k')$th element of $R$, $k, k' = 1, \ldots, m$, which implies a tri-diagonal sparse inverse correlation matrix. Supplementary Figure S.1 shows the graphs implied by the nonzero pattern of $R^{-1}$ used in the simulation study.

### 4.2. Mixed Gaussian, Binary and Ordinal Responses

We tested the proposed model in two different scenarios. Both scenarios consist of $m = 6$ responses with three Gaussian, one binary and two ordinal (with three and four categories) variables. In Scenario II, we generated 20 replicates with 100 samples and $p_k = 100$ predictors, $k = 1, \ldots, m$, whereas in Scenario I we simulated the same number of replicates with $n = 50$ and $p_k = 30$. We constructed the sparse vector of regression coefficients as described in Section 4.1 by setting the values the parameters $\pi_1, \pi_2, b$ and $s^2$ as shown in Table 1. With this choice of the parameters $\pi_1$ and $\pi_2$, in each scenario we simulated on average between four and five predictors associated with each response and with a small probability that they will be the same across responses since $\pi_1m < 1$.

To simulate realizations from the correlated responses, for $k = 1, 2, 3$, we set $F_k(x_{ik}; \beta_k, \theta_k)$ to be the cdf of the Gaussian distribution with mean $x_{ik}^T \beta_k$ and variance $\theta_k = 3$. For $k = 4$, we set $F_k(x_{ik}; \beta_k, \theta_k)$ to be the cdf of the Bernoulli distribution with mean $\Phi(x_{ik}^T \beta_k)$ and $\theta_k = 0$ and for $k = 5, 6$, we set $F_k(x_{ik}; \beta_k, \theta_k) = \Phi(\theta_{k4} - x_{ik}^T \beta_k)$ in order to simulate ordinal responses with $C_5 = 3$ and $C_6 = 4$ categories and cut-points $\theta_{k4}, c = 1, \ldots, C_k - 1$, which are drawn from a Unif(0, 1) and Unif(1, 2), respectively.

We used the MCMC sampler presented in Algorithm 1 to obtain posterior samples of the parameters and the latent variables of the BVSGCR model as well as to compute the MPPI for each predictor-response association. To estimate the parameters and the corresponding MPPIs for the single-response regression models, we employed widely-used MCMC algorithms for sparse linear Gaussian (George and McCulloch 1997) and non-linear (Holmes and Held 2006) regression models with the same proposal distribution for the selection of important predictors as described in Section 3.1. In all MCMC algorithms we chose the hyper-parameters $a_k$ and $b_k$ in Equation (5) following Kohn, Smith, and Chan (2001), where the mean and the variance of the beta distribution are matched with the a priori expected number of important predictors associated with each response (Err$y_{i \sim} = 5$) and its variance ($\text{var}(y_{i \sim}) = 9$). We also set $v = 1$ in Equation (4) since all predictors have been simulated with the same unit marginal variance. Finally, Supplementary Section S.1 presents the prior distribution on the response-specific parameters $\theta_k$ in the Gaussian and ordinal case. We ran each MCMC algorithm for 30,000 iterations using the first 10,000 as burn-in period, storing the outcome every 20 iterations to obtain 1,000 posterior samples for each model.

Table 2 displays the average area under the ROC curve (standard errors in brackets) for both Scenario I and II whereas Supplementary Figure S.2 presents the average (over 20 replicates) ROC curves for each one of the $m = 6$ responses simulated in Scenario II. Taken together, they indicate that for the analysis of correlated Gaussian, binary and ordinal responses, the BVSGCR model achieves a higher sensitivity in the Gaussian and the ordinal variables, and for the latter irrespectively of the number of the simulated categories. Similarly to the sparse SUR model with covariance selection, this is due to the ability of our model to account for the correlation between the responses which can induce false positive results when they are analyzed only

| Response | n | m | $\rho_k$ | $\pi_1$ | $\pi_2$ | b | $\sigma^2$ |
|----------|---|---|---------|--------|--------|---|--------|
| Scenario I & III Gaussian | 50 | 6 | 3 & 4 | 1 | 0.15 | 0.95 | 1 | 0.2 |
| Discrete | | | | | | | |
| Scenario II & IV Gaussian | 100 | 6 | 3 & 4 | 1 | 100 | 0.05 | 0.95 | 0.5 | 0.2 |
| Discrete | | | | | | | |

Table 2. Area under the ROC curves for the BVSGCR model and for independent single-response regression models in the simulated Scenario I and II. Results are averaged over 20 replicates with standard errors in brackets. Within each response, the best performance is highlighted in bold.
marginally. In addition, the proposal distribution for the non-zero regression coefficients in Equation (14) is tailored to take advantage of the estimated sparse inverse correlation structure, resulting in a more efficient algorithm for BVS.

For the binary response, the performance of the BVSGCR model and single-univariate regression model is almost identical. This is in keeping with Chib and Greenberg (1998) and Talhouk, Doucet, and Murphy (2012) that the estimates of the regression coefficients in multi-response Probit regression models are robust to the specification of the correlation structure including the case \( R = I_m \) which corresponds, in our framework, to the single-response Probit model.

We also investigated the effect of the “implicit marginalisation” of the regression coefficients in Equation (16) when the joint update of \((\beta_k, \gamma_k)\) in the M-H step is performed. To do so, we used the proposal density in Equation (15) that does not account for the correlation between responses and, more importantly, does not allow for the marginalization of the regression coefficients in the M-H step when Gaussian, binary and ordinal marginal distributions are jointly considered. Supplementary Figure S.4 shows that a better performance is achieved across all responses and in particular in the Gaussian and ordinal case when the proposal density in Equation (14) is used.

To assess the effect of the covariance selection procedure, we present in Supplementary Figure S.6, the ROC curves obtained by a specialized version of the proposed algorithm that does not allow any element of the inverse correlation matrix to be identically zero. Interestingly, the displayed ROC curves suggest that the Gaussian graphical model for covariance selection is crucial for an efficient identification of the important predictors. In particular, for the subset of Gaussian responses, the BVSGCR model with full \( R^{-1} \) is preferable than single-response linear regression models but it is not better than a model with sparse \( R^{-1} \). More important, in the case of discrete data, single-response regression models perform better in variable selection than the BVSGCR model when \( R^{-1} \) is a full matrix. A closer inspection of the MCMC output reveals that the selection of important predictors is affected by the difficult estimation of the inverse correlation matrix when the sample size is small and a full \( R^{-1} \) is enforced. With a larger sample size (\( n = 1000 \) data not shown) results are less affected by the specification of the covariance structure.

Finally, we also evaluated the estimation of the regression coefficients obtained by the BVSGCR model and compared with single-response regression models by using a modified version of the scoring rule presented in Gneiting and Raftery (2007). In our set-up, the interval score rewards narrow posterior credible intervals and incurs a penalty proportional to the significance level of the interval if the simulated nonzero regression coefficient is not included, see also Supplementary Section S.7 for its formal definition. Figure 1 displays the boxplots (over 20 replicates) of the average interval scores for the 95% credible intervals of the non-zero simulated regression coefficients in Scenario II for the BVSGCR model and single-response regression models. It is apparent that by using the proposed model, we obtained a more accurate estimation of the non-zero regression coefficients for all the responses, except, unsurprisingly, for the binary case.

4.3. Mixed Gaussian and Count Responses

In this section, we present the results of the application of the BVSGCR model in a simulated experiment in which the responses consist of a combination of one Gaussian and three count responses. We followed the same strategy described in Section 4.1 to generate the set of correlated predictors and the sparse vector of regression coefficients by choosing the parameters \( \pi_1, \pi_2, b \) and \( s^2 \) as described in Table 1 for the Gaussian and discrete responses. We also used the same correlation matrix \( R \) as described in Section 4.1.

We considered two different scenarios. Both consist of \( m = 4 \) responses with one Gaussian, two negative-binomial and one
binomial. In Scenario IV, we generated 20 replicates with \( n = 100 \) samples and \( p_k = 100 \) predictors, \( k = 1, \ldots, m \), whereas in Scenario III we simulated the same number of replicates with \( n = 50 \) and \( p_k = 30 \). Using Equation (3), we simulated the responses by specifying \( F_1(\cdot; x_{i1}, \beta_1, \theta_1) \) to be the cdf of the Gaussian distribution with mean \( x_{i1}^T \beta_1 \) and variance \( \theta_1 = 1 \), for \( k = 2 \) and \( k = 3 \), \( F_k(\cdot; x_{ik}, \beta_k, \theta_k) \) to be the cdf of the negative binomial distribution with mean \( \theta_k(1 + \exp(x_{ik}^T \beta_k))^{-1} \) with \( \theta_2 = \theta_3 = 0.5 \), and finally \( F_4(\cdot; x_{i4}, \beta_4, \theta_4) \) to be the cdf of the binomial distribution \( \text{Bin}(10, (1 + \exp(x_{i4}^T \beta_4))^{-1}) \) with \( \theta_4 = 10 \).

To estimate the parameters and the Gaussian latent variables of the BVSGCR model, we used the MCMC presented in Algorithm 1. We compared the results with the same MCMC algorithm used in Section 4.2 for the Gaussian response. For the count responses, we used the MCMC algorithm based on the representation of the negative binomial and binomial logistic regression model as a Gaussian regression model in auxiliary variables implemented in the R-package `pogit` by Dvorzak and Wagner (2016). After setting \( \text{E}(y_k) = 5 \) and \( \text{Var}(y_k) = 9 \), the parameters \( a_k \) and \( b_k \) of the beta prior on the probability of predictor-response association were chosen as in Section 4.2 and we matched the moments of `pogit` prior specification for \( y_k \) with these values. Finally, Supplementary Section S.1 presents the prior distributions on the response-specific parameters \( \theta_k \) for the Gaussian and negative-binomial responses. We ran the MCMC algorithms for 30,000 iterations, storing the outcome every 20 iterations, after a burn-in period of 10,000 iterations to obtain 1,000 posterior samples for each model.

Table 3 displays the average area under the ROC curve for both Scenario III and IV, whereas Supplementary Figure S.3 presents the average (over 20 replicates) ROC curves for each one of the \( m = 4 \) simulated responses in Scenario IV. Overall, it is evident, especially in the simulated Scenario IV, that by employing the proposed model, we achieved a better selection of important predictors compared to single-response regression models, apart from the binomial response. Similarly to the binary case, it seems there isn’t a clear advantage of the BVSGCR model over single-response nonlinear regression models when the marginal distribution is only parameterized by the probability of success.

Figure 2 displays the boxplot (over 20 replicates) of the average interval scores for the 95% credible intervals of the non-zero simulated regression coefficients in Scenario IV for the BVSGCR model and single-response regression models. The boxplots indicate that our model delivers more accurate estimates of the simulated regression coefficients than those obtained by using single-response regression models across all responses, including the binomial case. Thus, while the ROC curve for the binomial response shows that the ranking of the predictors based on the estimated MPPI is the same between the proposed model and the `pogit` algorithm, BVSGCR attains on average narrower 95% credible intervals.

---

**Table 3.** Area under the ROC curves for the BVSGCR model and for independent single-response regression models in the simulated Scenario III and IV. Results are averaged over 20 replicates with standard errors in brackets. Within each response, the best performance is highlighted in bold.

| Response | Regression model | Scenario III | Scenario IV |
|----------|-----------------|--------------|-------------|
|          |                 | \( n = 50 \) & | \( n = 100 \) & |
|          | \( p_k = 30 \)  | \( p_k = 100 \)  |
| Gaussian | BVSGCR          | \( 0.82 \) (0.13) | \( 0.86 \) (0.10) |
|          | Single-response | \( 0.80 \) (0.13) | \( 0.80 \) (0.12) |
| Negative Binomial | BVSGCR | \( 0.70 \) (0.16) | \( 0.76 \) (0.11) |
|          | Single-response | \( 0.68 \) (0.15) | \( 0.72 \) (0.11) |
| Negative Binomial | BVSGCR | \( 0.72 \) (0.14) | \( 0.76 \) (0.10) |
|          | Single-response | \( 0.70 \) (0.15) | \( 0.71 \) (0.13) |
| Binomial | BVSGCR          | \( 0.91 \) (0.10) | \( 0.91 \) (0.07) |
|          | Single-response | \( 0.91 \) (0.09) | \( 0.91 \) (0.06) |

---

**Figure 2.** Boxplot (over 20 replicates) of the average interval scores for the 95% credible interval of the non-zero simulated regression coefficients obtained by the BVSGCR model and by single-response regression models in the simulated Scenario IV.
For the simulated Scenario IV, Supplementary Figure S.7 compares the performance of the BVSGCR model with its specialized version when $R^{-1}$ is a full matrix. Interestingly, the performance is almost identical and both are better than single-response regression models, except for the binomial case. A closer look at the MCMC output reveals that, despite a sparse simulated inverse correlation structure and a small sample size ($n = 100$), the estimation of a full $R^{-1}$ is still feasible in this scenario with four simulated responses.

There is also another possible explanation regarding these results which is apparent in Supplementary Figure S.5. With the exception of the Gaussian outcome, for the binomial and negative-binomial responses, there is no “implicit marginalisation” of the regression coefficients in the M-H step. In this case, the proposal distribution in Equation (14) that takes into account the correlation between the responses performs no better than the proposal distribution in Equation (15) that does not make use of this information. It turns out that, when the “implicit marginalisation” is not possible, the acceptance probability of the joint update of $(\beta_k, y_k)$ in the M-H step seems to not take advantage of how accurately $R^{-1}$ is estimated and used in the proposal distribution.

5. Real Data Applications

We illustrate the features of the proposed BVSGCR model by applying it to two real datasets, Ataxia-Telangiectasia disorder and individuals suffering from Temporal Lobe Epilepsy, which are typical examples of data routinely collected in clinical research where a combination of discrete and/or continuous outcome variables are used to assess patients’ prognosis and disease progression. In the analysis of both real datasets our aim is twofold: (i) the identification of important associations between the outcome variables and the predictors that are either “response-specific” or “shared”, that is, predictors that are linked with several responses at the same time and (ii) the estimation the correlation pattern between the responses in order to shed light on their conditional dependence not explained by set of predictors considered. In the analysis of both real datasets our aim is twofold: (i) the identification of important associations between the outcome variables and the predictors. Missing values (completely at random or linked to some characteristic of the observed data. To overcome this problem, missing values in the outcome variables were imputed by modifying the designed MCMC presented in Algorithm 1 as suggested by Zhang et al. (2015) (see Supplementary Section S.6). Missing values in the predictors were imputed using the median of the observed values for each variable. To produce the results presented in this section, we ran the MCMC algorithms for 70,000 iterations. We considered the first 20,000 as burn-in and then we stored the output every 50th iteration in order to obtain 1,000 (thinned) samples from the posterior distributions of interest.

5.2. Ataxia-Telangiectasia Disorder

5.2.1. Data and BVSGCR Model

We applied our model on a dataset containing the measurements of 46 individuals suffering from Ataxia-Telangiectasia (A-T) disorder. A-T is a rare neurodegenerative disorder induced by mutations in the ATM gene. Our dataset is a subset of a larger multicentric cohort of 57 patients presented in Schon et al. (2019).

The dataset includes nine neurological responses (% missing values in brackets) and 13 predictors. In particular, four responses are continuous variables named as: Scale for Assessment and Rating of Ataxia (SARA) score (26%), Ataxia-Telangiectasia Neurological Examination Scale Toolkit (A-T NEST) score (17%), Age at first Wheelchair use (0%) and Alpha-Fetoprotein (AFP) levels (28%). Two responses are binary variables indicating the presence of Malignancy (0%) and the presence of Peripheral Neuropathy (11%) which is a term for a group of conditions in which the peripheral nervous system is damaged. Finally, three responses are ordinal variables which measure the overall Severity of the disorder (2%), its Progression (2%) and Eye Movements of the patients (2%). The set of predictors includes genetic (Genetic Group, Missense Mutation, that is, single base mutation responsible for the production of a different amino acid from the usual one, number of Mild Mutations, ATM Protein levels and Chromosomal Radiosensitivity, that is, whether X-ray exposure induces chromosomal aberrations in individuals with A-T) and immunological (immunoglobulin IgM, IgG2, IgG, IgA and IgE and immune CD4 and CD8 T-cell counts and CD19 B-cell counts) characteristics of the patients. In addition, we used an intercept term (with a diffuse normal prior centered in zero) and three confounders (age, gender and age of onset) always included in the regression model. Both confounders and predictors were standardized and the continuous variables quantile-transformed before the analysis.

We modeled the responses by specifying the BVSGCR model as follows. For each $i = 1, \ldots, n$ we set: for $k = 1, \ldots, 4$, $F_k(:, x_{ik}, \beta_k, \theta_k)$ to be the cdf of the Gaussian distribution with mean $x_{ik}^T \beta_k$ and variance $\theta_k$; for $k = 5, 6$, $F_k(:, x_{ik}, \beta_k, \theta_k)$ to be the cdf of the Bernoulli distribution with probability of success $\Phi(x_{ik}^T \beta_k)$ and $\theta_k = 0$; for $k = 7, 8, 9$, $F_k(:, x_{ik}, \beta_k, \theta_k) = \Phi(\theta_k c - x_{ik}^T \beta_k)$, where $\theta_k = (\theta_{k0}, \ldots, \theta_{k, C_k - 1})$ denotes the cut-points of the ordinal responses with $C_7 = C_9 = 3$ and $C_8 = 4$ categories, respectively. We used the prior distributions presented in Section 2.3 and in Supplementary Section S.1. Finally, we set $E(y_{ik}) = 3$ and $\text{var}(y_{ik}) = 2$ for all $k$. These values imply a priori a range of associations for each response between 0 and 8.
5.2. Results

Figure 3 displays the estimated MPPI for each predictor-response pair. Despite the small sample size, strong associations are detected in SARA score, A-T NEST score, AFP levels, Peripheral Neuropathy and Eye Movements and some evidence of association in Malignancy, Severity and Progression. Amongst the genetic predictors, Missense Mutation seems to play an important role in predicting the disease status and its progression (Eye Movements, Severity and Progression) as well as Malignancy and Peripheral Neuropathy (Schon et al. 2019). The number of Mild Mutations appears to influence A-T NEST score and AFP levels. Regarding the role of the immune system, it is well documented that patients suffering A-T have often a weakened defence mechanism. We confirm this clinical finding and, in contrast to the genetic risk factors, immunological predictors seem to be more “response-specific.”

Given the small sample size and (potentially important) unmeasured covariates, we do not expect that the genetic and immunological predictors are able to explain entirely the variability of the responses and their covariation. Therefore, it is important to model any source of extra variability that may induce false positives associations. Figure 4 shows the conditional dependence structure of the responses estimated by the BVSGCR algorithm with a decomposable model specification. Disease status and its progression (Severity and Progression) are closely linked with A-T NEST score and Age at first Wheelchair use, the latter also strongly related. Interestingly, Severity and Progression seem to capture different aspects of the disease since they are almost conditionally independent once the effect of Missense Mutation is accounted for (see Figure 3). A-T NEST score is also important in predicting the level of SARA score and Eye Movements. Finally, SARA score seems to be a good proxy for Peripheral Neuropathy.
We have also checked whether the assumption of decomposability is supported by the data. When a non-decomposable graphical model is specified, we found that the number of edges with a nonzero posterior probability of inclusion is higher than in the decomposable case, although there is a good agreement between the two instances of the BVSGCR algorithm regarding the most important edges, see Supplementary Figure S.13. Interestingly, the posterior mass is equally split between the two graphical models’ specifications, with a small advantage for the nondecomposable case (53%). The full results are reported in Supplementary Section S.8.7.

Table 4 presents, for each response, the estimated difference in the expected log-pointwise predictive density (ELPPD) between the BVSGCR model and independent single-response regression models. A positive difference indicates that the proposed model has better predictive performance (standard errors in brackets).

| Type                | Response       | Difference in ELPPD |
|---------------------|----------------|---------------------|
| Gaussian            | SARA           | 23.3 (6.4)          |
| Gaussian            | A-T NEST       | 23.0 (6.7)          |
| Gaussian            | Age Wheelchair | 25.9 (7.1)          |
| Gaussian            | Alpha-Fetoprotein | 19.4 (5.4)  |
| Binary              | Peripheral Neuropathy | 4.1 (1.8)  |
| Binary              | Malignancy     | 0.1 (0.6)           |
| Ordinal (3 categories) | Severity     | 17.1 (4.5)          |
| Ordinal (4 categories) | Progression   | 35.6 (7.6)          |
| Ordinal (3 categories) | Eye Movements | 15.1 (2.7)          |

Table 4. Difference in the expected log-pointwise predictive density (ELPPD) between the BVSGCR model and independent single-response regression models.

5.3. Patients with Temporal Lobe Epilepsy

5.3.1. Data and BVSGCR Model

We also applied the BVSGCR model to a second dataset consisting of $m = 5$ responses and $p_k = 162$, $k = 1, \ldots , m$, common predictors. We investigated the relationship between human cognition and epilepsy based on recent data collected by Johnson et al. (2016). The authors used $n = 122$ fresh-frozen whole-hippocampus samples, surgically resected from patients with Temporal Lobe Epilepsy (TLE) in order to determine whether genes belonging to their inferred gene-regulatory networks are related with human memory abilities and the number of seizures measured on the same individuals. More precisely, the responses (% of missing values in brackets) comprise the average number of self-reported daily Seizures for each patient (10% as we excluded some extremely large observations likely due to errors in self-reported number of Seizures), the memory category in which the patients have been assigned after the assessment by a neurologist (14%) and the results (Learning (15%), Post-Interference (15%) and Delayed Recall (15%)) of the Verbal Learning Test (Thiel et al. 2016) that quantifies the human cognition abilities. We also considered five confounding predictors: sex, age of manifestation of epilepsy, age at neurological assessment, anti-epileptic drugs load, handedness and laterality (brain lobe) of TLE. The 162 correlated genes were obtained from a network analysis described in Johnson et al. (2016). Both confounders and gene expression predictors were standardized to have unit variance.

We modeled the number of self-reported seizures with a negative binomial distribution and we used an ordinal variable for the memory categories in which the patients have been assigned by the neurologist. Finally, we assumed that the number of correct words that each patient recalls in each one of the three tasks of the Verbal Learning Test is distributed as a binomial random variable with 15 trials. To set the hyperparameters of the prior that controls the level of sparsity, we followed the same procedure used in the simulation study with $E(\gamma_k) = 5$ and $\text{var}(\gamma_k) = 9$ for all $k$. The prior distributions on $\theta_k$ for the negative binomial and the ordered categorical responses are presented in Section 2.3.

5.3.2. Results

Figure 5 displays the estimated MPPI of the associations between the 162 correlated genes and cognition abilities, the number of seizures and memory classification. The most striking finding is the ubiquitous role of RBFOX1 gene in Learning, Post-Interference and Delayed Recall as well as in Memory Category. It has been shown that mutations in this gene lead to neurodevelopmental disorder and it has also recently implied in cognitive functions (Davies et al. 2018). Regarding the Learning task, animal model studies have revealed critical functions for GABRB1 gene for maintenance of functioning circuits in the adult brain (Gehman et al. 2011). The genetic regulation of Delayed Recall is more complex and related to the difficult memory task that individuals were asked to perform.

In contrast to cognition abilities, the associations with the number of Seizures are less strong. This phenomenon can be explained by the quality of the self-reported data: before the analysis, we removed 10 measurements that appeared to be outliers. Despite that, the association with WNT3 gene seems interesting since deregulation in WNT signalling has a fundamental role in the origin of neurological diseases (Oliva, Vargas, and Inestrosa 2013).

We conclude the description of the association results by comparing the outcome of the proposed BVSGCR model with single-response regression models, see Supplementary Figure S.9. To conduct BVS for the ordinal response, we used the method proposed by Holmes, Denison, and Mallick (2002) and for negative binomial and binomial responses we utilized the auxiliary mixture sampling method of Frühwirth-Schnatter et al. (2009) implemented in the R-package pogsit (Dvorak and Wagner 2016). For the latter, we matched the moments of the prior on $\rho_k$ with the hyper-parameters of the BVSGCR sparsity prior. From the comparisons, we noticed that for the binomial responses the number of associations identified by the single-response regression model is either too large (Learning and Delayed Recall) or almost nil (Post-Interference). This may
Figure 5. Detection of important genes that predict human cognition abilities, number of Seizure and memory classification in 122 individuals with TLE. Marginal posterior probabilities of inclusion (MPPI) measures the strength of the predictor-response association. For each response, only associations with MPPI $> 0.05$ are highlighted.

Figure 6. Graph implied by the nonzero pattern of $R^{-1}$ in the TLE dataset and estimated by using the edge posterior probabilities of inclusion (EPPI) obtained by specifying a decomposable graphical model. Gray scale and edge thickness specify different levels of the EPPI (black thick line indicates large EPPI values).

depend on the Gibbs sampling search algorithm implemented in the R-package `pogit` that does not perform well when a large number of correlated predictors are considered (Bottolo and Richardson 2010). In contrast, our proposal distribution for $\gamma_k$ allows the quick detection of relevant predictors that explain a large fraction of the responses' variability, see Supplementary Figure S.11.

Figure 6 presents the conditional independence graph for the group of responses considered when a decomposable graphical model is assumed. Similarly to the A-T disorder, we do not expect to capture the whole responses' variability and their covariation given the set of predictors considered and (potentially important) unmeasured covariates. Interestingly, the responses of the Verbal Learning Test are all connected, with Memory Category linked only with Delayed Recall, suggesting that the neurologist's patients classification strongly reflects the Delayed Recall score. Finally, the number of Seizures is conditionally independent of the memory tasks of the Verbal Learning Test. This can be explained either by the self-reported quality of the data or by the fact that we removed the effect of age.
of onset which is known to be negatively correlated with both Seizures and cognition abilities.

We also checked if the assumption of decomposability of the graphical model is realistic. The support for nondecomposable graphs is overwhelmingly with 96% of the posterior mass concentrated on it. There is a general agreement regarding the detection of important edges between the two instances of the BVSGCR model, although for nondecomposable graphs, neurologist’s memory classification of the patients seems to be based, more reasonably, on the whole set of results of the Verbal Learning Test, see Supplementary Figure S.15.

Finally, Table 5 presents, for each response, the estimated difference in the expected log-pointwise predictive density (Vehtari, Gelman, and Gabry 2017) between the BVSGCR model and single-response regression models. They are all largely positive, indicating that the proposed model has better predictive performance, apart from the number of Seizures which is weakly associated with the set of genes and conditionally independent from the other responses. In any case, the small difference and the large standard deviation make it difficult to draw any clear conclusion about the best predictive model for this trait.

6. Discussion

In this article, we have presented a novel approach for BVS when a combination of discrete and/or continuous responses are jointly considered. The proposed method allows the exploration of the model space consisting of all possible subsets of predictors while estimating the conditional dependence structure among the responses and vice versa.

We have shown that for some continuous and discrete outcomes, including Gaussian, binary and ordinal responses, the regression coefficients can be “implicitly marginalized” in the acceptance probability of the M-H step for the joint update of \((\hat{\rho}_k, \gamma_k)\). This allows the reduction of the posterior correlation between these two quantities and thus improves the mixing of the designed sampler. The “implicit marginalisation” of the non-zero regression coefficients holds also when an unordered categorical variable is considered. In contrast to binary or ordinal responses, where only one latent variable is required, in the unordered categorical case, the state-space is expanded by \(C - 1\) latent variables where \(C\) is the number of categories. Combined with an effective proposal distribution for the latent binary vector, based on the marginal screening of important predictors for each response, our approach allows an efficient exploration of the ultra-high predictors model space.

We tested the proposed method on simulated and real datasets and compared it with widely used sparse Bayesian single-response linear and nonlinear regression models. In all examples considered, BVSGCR outperformed existing BVS algorithms in terms of selection of important predictors and/or estimation of the nonzero regression coefficients, except for the binary case. In the simulation study, we have also demonstrated that covariance selection is key when the sample size is small and the number of responses is large.

We conclude with some final remarks regarding directions for future research. As we have demonstrated in the simulation study, in the case of multiple-response count data, the proposed BVSGCR method performs better than single-response regression models because it exploits the Gaussian copula model. However, in the same simulated example, we have shown that the proposal distribution that takes into account the correlation between responses performs no better than the proposal distribution that does not use this information. When the “implicit marginalisation” is not possible, the M-H step doesn’t seem to take advantage of how accurately the residual correlations between the responses are estimated and this information used in the proposal distribution. Thus, it is paramount to specify the cdf of the marginal distribution so that it allows the marginalization of the regression coefficients. For count data, this may be accomplished by using the generalized ordered-response Probit (GORP) model presented in Castro, Paleti, and Bhat (2012) which can be expressed as a function of the Gaussian cdf. An interesting venue for future work will be the assessment of the similarities of the GORP model with Generalized Linear Models to fully exploit the benefits of the marginalization of the regression coefficients and the estimation of the residual correlations between the responses in the analysis of multivariate count data.

In summary, our new BVSGCR algorithm is tailored to jointly analyze correlated responses of diverse types with missing observations and a large set of predictors. Besides the application to clinically relevant data, the proposed method can be also used in the analysis of other problems for which sparse regression algorithms for combinations of discrete and/or continuous responses are required.

Supplemental Materials

Supplementary material: It provides additional material regarding the proposal density and the “implicit marginalization” of the regression coefficients as well as further results of the simulation study and the analysis of the real datasets.

R-package BVS4GCR: The package is freely available on https://github.com/b664/BVS4GCR. It includes examples that explain how to generate the simulated data and run the algorithm.

Acknowledgments

The authors are thankful to Katherine Schon and Michael Johnson for providing the data of the Ataxia-Telangiectasia disorder and Temporal Lobe Epilepsy examples and to Hélène Ruffieux and Verena Zuber for insightful comments. The authors are also grateful to the editor, associate editor and two anonymous referees for their valuable comments that greatly improved the presentation of the article.
Song, P. X.-K. (2000), “Multivariate Dispersion Models Generated from Gaussian Copula,” Scandinavian Journal of Statistics, 27, 305–320. [579,580]

Song, P. X.-K., Li, M., and Yuan, Y. (2009), “Joint Regression Analysis of Correlated Data Using Gaussian Copulas,” Biometrics, 65, 60–68. [578,580]

Talhouk, A., A. Doucet, and K. Murphy (2012), “Efficient Bayesian Inference for Multivariate Probit Models with Sparse Inverse Correlation Matrices,” Journal of Computational and Graphical Statistics, 21, 739–757. [579,580,581,583,585]

Thiel, C. M., Özyurt, J., Nogueira, W., and Puschmann, S. (2016), “Effects of Age on Long Term Memory for Degraded Speech,” Frontiers in Human Neuroscience, 10, 473. [589]

Van Dyk, D. A., and Meng, X.-L. (2001), “The Art of Data Augmentation,” Journal of Computational and Graphical Statistics, 10, 1–50. [579]

Vehtari, A., Gabry, J., Yao, Y., and Gelman, A. (2018), “LOO: Efficient Leave-one-out Cross-validation and WAIC for Bayesian Models. R package version 2.0.0.” [587]

Vehtari, A., A. Gelman, and Gabry, J. (2017), “Practical Bayesian Model Evaluation Using Leave-one-out Cross-validation and WAIC,” Statistical Computing, 27, 1413–1432. [587,589,591]

Wang, H. (2010), “Sparse Seemingly Unrelated Regression Modelling: Applications in Finance and Econometrics,” Computational Statistics and Data Analysis, 54, 2866–2877. [578,580]

Wang, H., and Li, S. Z. (2012), “Efficient Gaussian Graphical Model Determination Under G-Wishart Prior Distributions,” The Electronic Journal of Statistics, 6, 168–198. [583]

Webb, E. L., and Forster, J. J. (2008), “Bayesian Model Determination for Multivariate Ordinal and Binary Data,” Computational Statistics and Data Analysis, 52, 2632–2649. [579,580]

Yu, Y. and Meng, X.-L. (2011), “To Center or Not to Center: That is Not the Question–An Ancillarity–Sufficiency Interweaving Strategy (ASIS) for Boosting MCMC Efficiency,” Journal of Computational and Graphical Statistics, 20, 531–570. [582]

Zellner, A. (1962), “An Efficient Method of Estimating Seemingly Unrelated Regressions and Tests for Aggregation Bias,” Journal of American Statistical Association, 57, 348–368. [578]

Zhang, X., Boscardin, W. J., Belin, T. R., Wan, X., He, Y., and Zhang, K. (2015), “A Bayesian Method for Analyzing Combinations of Continuous, Ordinal, and Nominal Categorical Data with Missing Values,” Journal of Multivariate Analysis, 135, 43–58. [578,587]