Joint Models for the Association of Longitudinal Binary and Continuous Processes with Application to a Smoking Cessation Trial

Xuefeng Liu

Assistant Professor

Department of Internal Medicine

Wayne State University

Michael J. Daniels

Professor

Department of Statistics

University of Florida

Bess Marcus

Professor

Department of Psychiatry & Human Behavior

Brown University

Abstract

Joint models for the association of a longitudinal binary and a longitudinal continuous process are proposed for situations where their association is of direct interest. The models are parameterized such that the dependence between the two processes is characterized by unconstrained regression coefficients. Bayesian variable selection techniques are used to parsimoniously model these coefficients. An MCMC sampling algorithm is developed for sampling from the posterior distribution, using data augmentation steps to handle missing data. Several technical issues are addressed to implement the MCMC algorithm efficiently. The models are motivated by, and are used for, the analysis of a smoking cessation clinical trial in which an important question of interest was the effect of the (exercise) treatment on the relationship between smoking cessation and weight gain.

KEY WORDS: Joint Models, Stochastic Search Variable Selection, MCMC, Data Augmentation, Dependence, Parameter Expansion, Calibrated Posterior Predictive *p*-value.

1 Introduction

In some longitudinal studies, although one time-varying outcome may be of primary interest, several related processes are being measured. Some examples in smoking cessation studies include smoking status and weight change, and smoking status and alcohol use. In such studies, the association between the processes can reveal a great deal about the mechanism of behavior change. For example, a motivation for using exercise as an adjunct therapy for smoking cessation is to reduce the dependence between weight gain and relapse back to smoking and between the fear of weight gain and the inability to make a successful quit attempt. In this study we will build models in the setting of two processes: a longitudinal binary process (e.g., smoking cessation) and a longitudinal continuous process (e.g., weight change). Our primary interest is to model the association between the two processes. We will apply the models to a recent smoking cessation trial (Marcus et al., 2003) where the investigators were interested in the relation between smoking status and weight change.

Our approach builds on and extends recent research on joint modelling of mixed outcomes and Bayesian variable selection. For joint modeling, a well-known technique of joint modelling of mixed outcomes is based on introducing a partly observed random variable following a bivariate normal distribution, where one component defines the continuous outcome, while the second component is latent and defines the binary outcome through the common probit transformation. Papers taking this type of approach include Catalano and Ryan (1992), Cox and Wermuth (1992), Dunson (2000), Fitzmaurice and Laird (1995), Gueorguieva and Agresti (2001), Regan and Catalano (1999), Roy and Lin (2000) and Sammel *et al* (1997). In this paper, we shall extend this approach to longitudinal data with T individual measurements by considering a partly observed random variable following a $2T$-variate normal distribution where the first T components define the binary outcomes by applying probit transformations while the last T components are continuous outcomes. Moreover, the main question of interest in previous studies was the effect of some treatment or therapy on the mean of the response vector. The effect of the treatment on the association between two outcomes was not of main interest. We will propose similar models in which the association between two processes is of concern. To do this, we use a Bartlett decomposition of the covariance matrix (Bartlett, 1933).

The association matrix induced by the Bartlett decomposition is high-dimensional and expected to be sparse so we borrow ideas from the Bayesian variable selection literature to reduce the number of parameters (George and McCulloch, 1993, 1997; Smith and Kohn, 2002). Other related work includes Carlin and Chib (1995), Chipman (1996), Hoeting, Raftery and Madigan (1996), and Wakefield and Bennett (1996). George and McCulloch developed stochastic search variables selection (SSVS) to select promising subsets of a set of covariates $X_1, \cdots, X_p$ for further consideration in regression models. Smith and Kohn (2002) proposed similar techniques to model a covariance matrix with high dimension for longitudinal data. In this study, we construct a hierarchical prior to parsimoniously model the association between two longitudinal processes by extending the idea in Smith and Kohn (2002). The hierarchical specification has the advantage that potential zeros in the association matrix can be identified and estimates of the parameters can be calculated that account for the model uncertainty associated with determining which elements are zero.

We develop an MCMC algorithm to estimate the posterior distribution of the parameters in the models. To implement the MCMC algorithm more efficiently, we address several technical issues, including efficiently sampling from truncated multivariate normal (TMVN) distributions and efficiently sampling a correlation matrix from its full conditional distribution. Geweke (1991) and Robert (1995) proposed a Gibbs sampling algorithm to sample from a TMVN distribution. This algorithm is somewhat inefficient in that it requires repeatedly evaluating conditional means and variances from univariate conditional normals and can result in high autocorrelations in the chain. Barnard, McCulloch and Meng (2000), and Chib and Greenberg (1998) suggested using the Griddy Gibbs (GG) sampler and the random walk Metropolis-Hastings (RW-MH) algorithm, respectively, to sample a correlation matrix. Although the GG sampler is simple to implement, it is not computationally efficient. The RW-MH algorithm has the problem of potentially slow mixing. We shall propose better ways to handle these issues.

The paper is organized as follows. We introduce a smoking cessation clinical trial that motivates this research in Section 2. In Section 3, we propose joint models and hierarchical priors used to parsimoniously model the association between two processes. In Section 4, MCMC sampling techniques to estimate the posterior distribution of parameters will be described, and several technical issues regarding efficient implementation of the MCMC algorithm will be addressed. We discuss the deviance information criterion (DIC) for model comparison and the calibrated posterior predictive p-value for goodness of fit in Section 5. Finally, we present the results and conclusions from the analysis of the clinical trial in Section 6.

2 Application: Smoking Cessation Trial

Commit to Quit II (CTQ II) (Marcus *et al*, 2003; Marcus *et al*, 2005) was a 4-year randomized controlled clinical trial. It was designed to test the efficacy of moderate-intensity physical activity as an aid for smoking cessation among women. This study was a logical progression of previous work (CTQ I) (Marcus *et al*, 1997; Marcus *et al*, 1999) on the efficacy of vigorous-intensity exercise to aid smoking cessation and weight regulation in women smokers as moderate-intensity exercise is less arduous and can be performed by healthy individuals without medical supervision. In the CTQ II trial, 217 healthy women between the ages of 18-65 who had regularly smoked five or more cigarettes per day for at least 1 year and who had routinely participated in moderate or vigorous intensity physical activity for 90 minutes or less each week were recruited, and were randomized to one of the two conditions (treatments): 1) a moderate-intensity exercise condition; 2) a contact condition. We will refer to these two treatments as *exercise* and *wellness*, respectively in the rest of the manuscript. All recruited women participated in an 8-week cognitive-behavioral group-based smoking cessation program, followed by a 12-month follow-up. Participants in the moderate-intensity exercise condition were required to attend one supervised exercise session per week that occurred on their smoking cessation treatment night. They were also given written instructions for home exercises. The duration and intensity of the exercise were gradually increased to 165 minutes per week that could be performed on-site or at home. Participants in the contact condition were given lectures, films and handouts on a variety of health and lifestyle issues. All participants were called and encouraged to perform makeups if they failed to attend any session during the 8 weeks of treatments. Smoking status was determined via self-report and carbon monoxide testing at each session. Additionally, participants were weighed on a weekly basis during the 8 weeks of treatments. This design allowed for a comparison of the effect of moderate-intensity physical activity plus standard smoking cessation with the effect of contact plus standard smoking cessation.

The primary outcome was quit status (a longitudinal binary outcome), but another outcome, weight change (a longitudinal continuous outcome), was also measured. The investigators were interested in two questions: 1) Does moderate-intensity exercise have significant effects on smoking cessation? In the study of the CTQ I trial, the investigators found significant differences between vigorous activity and contact control group through 12 months of follow up; 2) Does exercise effect smoking cessation by weakening the association between smoking status and weight gain? The second question motivates our development of joint models for the association of longitudinal binary and continuous processes to examine the differential patterns of association across treatments in Section 3. However, our approach also allows us to answer question 1) as well.

3 Joint Models and Priors

3.1 Joint Models

Several authors have developed joint models for analysis of multivariate longitudinal data using latent normal variables (Daniels and Normand, 2006; Dunson, 2003; Gueorguieva and Sanacora, 2006). In this section, we propose similar joint models for the association of a longitudinal binary and a longitudinal continuous process. However, in our setting the time-dependent covariance matrix is modelled as a function of predictors and is of primary interest.

Denote the binary outcome (in our example, smoking status) for subject j in treatment i at week t $(i = 1, \cdots, m; j = 1, \cdots, n_i; t = 1, \cdots, T)$ by $Q_{ij,t}$ and the continuous outcome (in our example, weight change) by $W_{ij,t}$. Define the vector of responses for binary and continuous outcomes as $\boldsymbol{Q}_{ij} = (Q_{ij,1}, \cdots, Q_{ij,T})'$ and $\boldsymbol{W}_{ij} = (W_{ij,1}, \cdots, W_{ij,T})'$, respectively. We also define a vector of latent variables underlying the binary vector $\boldsymbol{Q}_{ij}$ to be $\boldsymbol{Y}_{ij} = (Y_{ij,1}, \cdots, Y_{ij,T})'$. Suppose that $\boldsymbol{V}_{ij}$ is a vector of joint processes such that $\boldsymbol{V}_{ij} = (\boldsymbol{Y}_{ij}', \boldsymbol{W}_{ij}')'$. Then the joint distribution of binary and continuous variables over time can be modelled using the multivariate normal specification $(\boldsymbol{Y}_{ij}', \boldsymbol{W}_{ij}')' \sim \mathrm{N}(\boldsymbol{X}_{ij}\boldsymbol{\beta}, \boldsymbol{\Sigma}_i)$,

$$\boldsymbol{V}_{ij} = \begin{pmatrix} \boldsymbol{Y}_{ij} \\ \boldsymbol{W}_{ij} \end{pmatrix} = \boldsymbol{X}_{ij}\boldsymbol{\beta} + \boldsymbol{\epsilon}_i, \tag{3.1}$$

where $\boldsymbol{X}_{ij}$ is the design matrix, $\boldsymbol{\beta}$ is the vector of regression coefficients, and $\boldsymbol{\epsilon}_i \sim \mathrm{N}(\boldsymbol{0}, \boldsymbol{\Sigma}_i)$ with $\boldsymbol{\Sigma}_i = \begin{pmatrix} \boldsymbol{\Sigma}_{i,11} & \boldsymbol{\Sigma}_{i,12} \\ \boldsymbol{\Sigma}_{i,21} & \boldsymbol{\Sigma}_{i,22} \end{pmatrix}$. Using the probit formulation for the binary process, we have $Q_{ij,t} = \mathrm{I}\{Y_{ij,t} > 0\}$. To estimate the association between $\boldsymbol{Q}_{ij}$ and $\boldsymbol{W}_{ij}$, we need models for $\boldsymbol{\Sigma}_{i,12}$ as a function of treatment (and/or other subject specific covariates that might affect this relationship). However, both $\boldsymbol{\Sigma}_{i,12}$ and the entire covariance matrix $\boldsymbol{\Sigma}_i$ are difficult to model due to positive definiteness constraints (Daniels and Kass, 2001; Pourahmadi and Daniels, 2002) in addition to being high-dimensional for each subject. To address this problem, we factor the joint distribution of $\boldsymbol{Y}_{ij}$ and $\boldsymbol{W}_{ij}$ into two components: a marginal model for $\boldsymbol{Y}_{ij}$ and a correlated regression model for $\boldsymbol{W}_{ij}$ given $\boldsymbol{Y}_{ij}$ by extending the ideas from Fitzmaurice and Laird (1995) and Gueorguieva and Agresti (2001). Let $\boldsymbol{X}_{ij} = \begin{pmatrix} \boldsymbol{X}_{1ij} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{X}_{2ij} \end{pmatrix}$ and $\boldsymbol{\beta} = \begin{pmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{pmatrix}$, then the new models can be expressed as

$$\boldsymbol{Y}_{ij} = \boldsymbol{X}_{1ij}\boldsymbol{\beta}_1 + \boldsymbol{\epsilon}_{1i} \tag{3.2}$$

$$\boldsymbol{W}_{ij} = \boldsymbol{X}_{2ij}\boldsymbol{\beta}_2 + \boldsymbol{B}_i(\boldsymbol{Y}_{ij} - \boldsymbol{X}_{1ij}\boldsymbol{\beta}_1) + \boldsymbol{\epsilon}_{2i}, \tag{3.3}$$

where $\boldsymbol{B}_i = \boldsymbol{\Sigma}_{i,21}\boldsymbol{\Sigma}_{i,11}^{-1}$ is the matrix that reflects association between $\boldsymbol{Y}_{ij}$ and $\boldsymbol{W}_{ij}$, $\boldsymbol{\epsilon}_{1i} \sim \mathrm{N}(\boldsymbol{0}, \boldsymbol{\Sigma}_{i,11})$, and $\boldsymbol{\epsilon}_{2i} \sim \mathrm{N}(\boldsymbol{0}, \boldsymbol{\Sigma}_{i,22}^*)$ with $\boldsymbol{\Sigma}_{i,22}^* = \boldsymbol{\Sigma}_{i,22} - \boldsymbol{\Sigma}_{i,21}\boldsymbol{\Sigma}_{i,11}^{-1}\boldsymbol{\Sigma}_{i,12}$. The reparameterization of $\boldsymbol{\Sigma}_i$ to $(\boldsymbol{\Sigma}_{i,11}, \boldsymbol{B}_i, \boldsymbol{\Sigma}_{i,22}^*)$ is known in the literature as the Bartlett decomposition of a covariance matrix (Bartlett, 1933).

It is easy to see that (3.2) is a correlated probit model and (3.3) is a standard correlated regression model, conditional on the latent variable $\boldsymbol{Y}_{ij}$. For identifiability, it is common to restrict $\boldsymbol{\Sigma}_{i,11}$ to be a correlation matrix $\boldsymbol{R}_{i,11}$ (Chib and Greenberg, 1998); in the rest of this paper, we will use $\boldsymbol{R}_{i,11}$ instead of $\boldsymbol{\Sigma}_{i,11}$ as notation for the covariance matrix in (3.2). The advantage of this factorization is that the components of $\boldsymbol{B}_i$ in (3.3) are directly related to the variance and correlation terms in $\boldsymbol{\Sigma}_{i,12}$. In addition, this factorization provides a convenient parameterization to examine the association between $\boldsymbol{Y}_{ij}$ ($\boldsymbol{Q}_{ij}$) and $\boldsymbol{W}_{ij}$ since the components of the $\boldsymbol{B}_i$ matrix are unconstrained.

Beside being unconstrained, the association matrix $\boldsymbol{B}_i$ in model (3.3) can be easily interpreted. The tth row of $\boldsymbol{B}_i$ reflects the association of the continuous process at week t with the binary process at all weeks ($t = 1, \cdots, T$). In particular, it corresponds to the regression model $w_{ij,t}|\boldsymbol{Y}_{ij} = \boldsymbol{x}'_{2ij,t}\boldsymbol{\beta}_2 + \boldsymbol{b}'_{i,t}(\boldsymbol{Y}_{ij} - \boldsymbol{X}_{1ij}\boldsymbol{\beta}_1) + \boldsymbol{\epsilon}_{2i,t}$, where $\boldsymbol{b}_{i,t} = (b_{i,t1}, \cdots, b_{i,tT})'$ is the tth row of $\boldsymbol{B}_i$. Since the covariates associated with $\boldsymbol{b}_{i,t}$, $\boldsymbol{Y}_{ij} - \boldsymbol{X}_{1ij}\boldsymbol{\beta}_1$, are centered with variance one (recall the marginal covariance matrix of $\boldsymbol{Y}_{ij}$ is a correlation matrix), the components of $\boldsymbol{B}_i$ are *standardized* regression coefficients. This property of the components of $\boldsymbol{B}_i$ will facilitate between-component comparisons and motivate ideas for modeling it parsimoniously.

3.2 Priors for Parameters in Joint Models

For Bayesian inference, we need to specify priors for parameters in the models in Section 3.1. Denote by $\boldsymbol{b}_i$ the column vector obtained by stringing the rows of $\boldsymbol{B}_i$ ($i = 1, \cdots, m$); that is, $\boldsymbol{b}_i = (B_{i,11}, \ldots, B_{i,1T}, \ldots, B_{i,TT})'$. Let $\boldsymbol{R}_1 = (\boldsymbol{R}'_{1,11}, \cdots, \boldsymbol{R}'_{m,11})'$, $\boldsymbol{b} = (\boldsymbol{b}'_1, \cdots, \boldsymbol{b}'_m)'$ and $\boldsymbol{\Sigma}^*_2 = (\boldsymbol{\Sigma}^{*'}_{1,22}, \cdots, \boldsymbol{\Sigma}^{*'}_{m,22})'$. We write the joint prior in our models as:

$$\pi(\boldsymbol{\beta}, \boldsymbol{R}_1, \boldsymbol{b}, \boldsymbol{\Sigma}^*_2) = \prod_{i=1}^{m} \pi(\boldsymbol{\beta})\pi(\boldsymbol{R}_{i,11})\pi(\boldsymbol{b}_i|\boldsymbol{\beta}, \boldsymbol{\Sigma}^*_{i,22})\pi(\boldsymbol{\Sigma}^*_{i,22}). \tag{3.4}$$

The above specification implies that 1) $\boldsymbol{\beta}$, $\boldsymbol{b}_i$ and $\boldsymbol{\Sigma}^*_{i,22}$ are *a priori* jointly independent of $\boldsymbol{R}_{i,11}$; 2) Marginally, $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}^*_{i,22}$ are *a priori* independent. Since we have little prior information for $\boldsymbol{\beta}$, $\boldsymbol{\Sigma}^*_{i,22}$ and $\boldsymbol{R}_{i,11}$, we specify flat priors for $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}^*_{i,22}$, and a joint uniform prior (derived by Barnard *et al* (2000)) for $\boldsymbol{R}_{i,11}$,

$$\pi(\boldsymbol{\beta}) \propto 1 \tag{3.5}$$

$$\pi(\boldsymbol{\Sigma}^*_{i,22}) \propto \mathrm{I}\{\boldsymbol{\Sigma}^*_{i,22} \in (-\infty, \infty)^{\frac{T(T+1)}{2}} \text{ and } \boldsymbol{\Sigma}^*_{i,22} \text{ is positive definite}\} \tag{3.6}$$

$$\pi(\boldsymbol{R}_{i,11}) \propto \mathrm{I}\{r_{i,jk} : r_{i,jk} = 1\ (j = k), |r_{i,jk}| < 1\ (j \neq k) \text{ and } \boldsymbol{R}_{i,11} \text{ is positive definite}\}, \tag{3.7}$$

where $r_{i,jk}$ $(j \neq k, j, k = 1, \cdots, T)$ is the off-diagonal element of the jth row and kth column in $\boldsymbol{R}_{i,11}$.

3.2.1 Prior for the Elements of the Association Matrix

We now provide details on the prior $\pi(\boldsymbol{b}_i|\boldsymbol{\beta}, \boldsymbol{\Sigma}^*_{i,22})$. Since the components of $\boldsymbol{b}_i$ are the regression coefficients of $\boldsymbol{W}_{ij}$ on $\boldsymbol{Y}_{ij}$, we expect many of the components of $\boldsymbol{b}_i$ to be zeros based on conditional independence (Markov type) arguments for longitudinal data. For example, consider the regression for w_{ijt}. Once we condition on $\{Y_{ij,k} : t-1 \leq k \leq t+1\}$ (current value and lag one value forward and backward), we might expect w_{ijt} is independent of $\{Y_{ij,k} : k > t+1, k < t-1\}$ (values more than lag one way). To incorporate these features into the model, we specify a hierarchical prior distribution that essentially allows components of $\boldsymbol{b}_i$ to be zeros, borrowing ideas from Smith and Kohn (2002). This will also facilitate reducing the number of dependence parameters, which can be quite large.

A key feature of the hierarchical prior is that each component of $\boldsymbol{b}_i$ (recall each component is on the same scale since they are standardized regression coefficients) can be exactly zero with positive probability. To attain this, we introduce the latent indicator vector $\boldsymbol{\delta}_i$ associated with $\boldsymbol{b}_i$ such that

$$b_{i,tl} = \begin{cases} 0 & \text{if } \delta_{i,tl} = 0 \\ 1 & \text{if } \delta_{i,tl} = 1 \end{cases},$$

where $b_{i,tl}$ is the component of the tth row and lth column in $\boldsymbol{B}_i$ and $\delta_{i,tl}$ is the corresponding binary indicator of $\boldsymbol{\delta}_i$ which is associated with $b_{i,tl}$. The nonzero components of the vector $\boldsymbol{b}_i$ (i.e., the components for which $\delta_{i,tl} = 1$) are given a normal prior (conditional on $\delta_{i,tl} = 1$) with mean

$$E[\boldsymbol{b}_i|\boldsymbol{Y}, \boldsymbol{\beta}, \boldsymbol{\delta}, \boldsymbol{\Sigma}^*_{22}] = \left(\sum_{j=1}^{n_i} \boldsymbol{C}'_{ij\delta}(n_i\boldsymbol{\Sigma}^*_{i,22})^{-1}\boldsymbol{C}_{ij\delta}\right)^{-1} \sum_{j=1}^{n_i} \boldsymbol{C}'_{ij\delta}(n_i\boldsymbol{\Sigma}^*_{i,22})^{-1}\boldsymbol{Z}_{ij}$$

and covariance matrix

$$\text{Var}[\boldsymbol{b}_i|\boldsymbol{Y}, \boldsymbol{\beta}, \boldsymbol{\delta}, \boldsymbol{\Sigma}^*_{22}] = \left(\sum_{j=1}^{n_i} \boldsymbol{C}'_{ij\delta}(n_i\boldsymbol{\Sigma}^*_{i,22})^{-1}\boldsymbol{C}_{ij\delta}\right)^{-1}.$$

where $\boldsymbol{Z}_{ij} = \boldsymbol{W}_{ij} - \boldsymbol{X}_{2ij}\boldsymbol{\beta}_2$, $\boldsymbol{C}_{ij\delta}$ is obtained by removing from $\boldsymbol{C}_{ij}$ the columns corresponding to zero elements of $\boldsymbol{b}_i$, and $\boldsymbol{C}_{ij}$ is $T \times T^2$ matrix with elements that are functions of $\boldsymbol{Y}_{ij}$ and the tth row $\boldsymbol{C}_{ij,t} = (\boldsymbol{0}'_{ij,1}, \cdots, \boldsymbol{0}'_{ij,(t-1)}, (\boldsymbol{Y}_{ij} - \boldsymbol{X}_{1ij}\boldsymbol{\beta}_1)'_t, \boldsymbol{0}'_{ij,(t+1)}, \cdots, \boldsymbol{0}'_{ij,T})$. The vector $\boldsymbol{\delta}_i$ is then given the following prior,

$$\pi(\boldsymbol{\delta}_i|p_i) = \prod_{t=1}^{T}\prod_{l=1}^{T} \pi(\delta_{i,tl}|p_i) = \prod_{t=1}^{T}\prod_{l=1}^{T} \text{Bernoulli}(p_i^{|t-l|^{1/a_0}+1}) \tag{3.8}$$

$$\pi(p_i) = \text{Beta}(r_i, \lambda_i), \tag{3.9}$$

where $\boldsymbol{\delta}_i = (\delta_{i,11}, \cdots, \delta_{i,1T}, \cdots, \delta_{i,T1}, \cdots, \delta_{i,TT})'$, $\delta_{i,tl}$ $(i = 1, \cdots, m;\ l, t = 1, \cdots, T)$ is the element of $\boldsymbol{\delta}_i$ associated with $b_{i,tl}$ which is the regression coefficient of $w_{ij,t}$ on $y_{ij,l}$, a_0 is a tuning parameter, and $(r_i,\ \lambda_i)$ are the corresponding hyper-parameters.

The prior for $\boldsymbol{b}_i$ given $\boldsymbol{\delta}_i$ was derived based on

$$\pi(\boldsymbol{b}_i|\boldsymbol{\delta}_i, \boldsymbol{\beta}, \boldsymbol{\Sigma}^*_{i,22}) \propto f^{\frac{1}{n_i}}(\boldsymbol{W}_i|\boldsymbol{Y}_i, \boldsymbol{\beta}, \boldsymbol{b}_i, \boldsymbol{\delta}_i, \boldsymbol{\Sigma}^*_{i,22}). \tag{3.10}$$

The rationale for (3.10) is that the prior only provides $1/n$th of the weight provided by the likelihood.

Based on this prior construction, the quantity $p_i^{|t-l|^{1/a_0}+1}$ in prior (3.8) can be thought of as the prior probability that $b_{i,tl}$ will require a non-0 estimate, and $|t-l|^{1/a_0}$ implies that $b_{i,tl}$ is more likely to be 0 as $|t-l|^{1/a_0}$ gets bigger; a_0 is a tuning parameter that controls the rate of decay of the probability of a non-zero component as a function of lag. (3.8) implies that the components of $\boldsymbol{B}_i$ become smaller *a priori* as they move away from the main diagonal (in the longitudinal setting, become smaller as moving farther away in time). This exponent can also be adjusted if we expect, *a priori*, a lagged relationship.

3.2.2 Using the Prior for the Association Matrix in Practice

A complication with the prior for $\boldsymbol{b}_i$ is that it is a function of $(\boldsymbol{Y}, \boldsymbol{\beta}, \boldsymbol{\Sigma}^*_{i,22})$, which will complicate the form of the full conditional distributions for the MCMC algorithm described in the web appendix, Section I (www.amstat.org/publications/ jasa/ supplemental materials). To address this issue, we replace the mean and covariance for the prior with $\hat{\boldsymbol{\mu}}_{b_{i\delta}}$ and covariance $\hat{\boldsymbol{\Sigma}}_{b_{i\delta}}$ defined below. These quantities are computed from an MCMC run with $\boldsymbol{b}_i = 0$ as $\hat{\boldsymbol{\mu}}_{b_{i\delta}} = \sum_k \mathrm{E}(\boldsymbol{b}_i|\boldsymbol{Y}^{(k)}, \boldsymbol{\beta}^{(k)}, \boldsymbol{\delta}, \boldsymbol{\Sigma}^{(k)*}_{22})$ and $\hat{\boldsymbol{\Sigma}}_{b_{i\delta}} = \sum_k \mathrm{Var}(\boldsymbol{b}_i|\boldsymbol{Y}^{(k)}, \boldsymbol{\beta}^{(k)}, \boldsymbol{\delta}, \boldsymbol{\Sigma}^{(k)*}_{22})$ where k indexes the posterior sample. An alternative would be to just use a mean and variance with the posterior means of $(\boldsymbol{Y}, \boldsymbol{\beta}, \boldsymbol{\Sigma}^*_{22})$ plugged in.

3.2.3 Propriety of the Joint Posterior Distribution

Given that we have little prior information on $\boldsymbol{\beta}$, $\boldsymbol{\Sigma}^*_{i,22}$ and $\boldsymbol{R}_{i,11}$, we proposed using flat priors for $\boldsymbol{\beta}$, $\boldsymbol{\Sigma}^*_{i,22}$ and $\boldsymbol{R}_{i,11}$ in Section 3.2. Because the priors for $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}^*_{i,22}$ are improper, we need to check that the joint posterior distribution is integrable. In Web appendix, Section II (www.amstat.org/publications/ jasa/ supplemental materials), we state (and prove) a theorem that guarantees integrability of the posterior when four easy to check conditions are met. The proof of this theorem extends some results in Chen and Shao (1999) for the propriety of posterior distributions for multivariate categorical response models and Daniels (2006) for the propriety of the posterior for linear regression with correlated and/or heterogeneous errors.

4 Posterior Sampling

4.1 The MCMC Sampling Algorithm

We develop an MCMC algorithm to sample from the posterior distribution of the parameter $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{R}_1, \boldsymbol{b}, \boldsymbol{\delta}, \boldsymbol{p}, \boldsymbol{\Sigma}^*_2)$, where $\boldsymbol{\delta} = (\boldsymbol{\delta}'_1, \cdots, \boldsymbol{\delta}'_m)'$, $\boldsymbol{p} = (p_1, \cdots, p_m)'$, and

$\boldsymbol{R}_1$, $\boldsymbol{b}$ and $\boldsymbol{\Sigma}_2^*$ are defined as in (3.4). To simplify implementation of the algorithm, we include a data augmentation (DA) (Tanner and Wong, 1987) step that will be used to impute the latent data and the missing values. Let $\boldsymbol{Q}_{obs}$ be the vector of observed binary outcomes, $\boldsymbol{Q}_{mis}$ be the vector of missing binary outcomes, $\boldsymbol{W}_{obs}$ be the vector of observed continuous outcomes and $\boldsymbol{W}_{mis}$ be the vector of missing continuous outcomes. Denote $\boldsymbol{Y}_{obs}$ to be the latent vector associated with $\boldsymbol{Q}_{obs}$ and $\boldsymbol{Y}_{mis}$ to be the latent vector related to $\boldsymbol{Q}_{mis}$. Define $\boldsymbol{Y}$ to be $(\boldsymbol{Y}_{obs}, \boldsymbol{Y}_{mis})$, $\boldsymbol{Q}$ to be $(\boldsymbol{Q}_{obs}, \boldsymbol{Q}_{mis})$ and $\boldsymbol{W}$ to be $(\boldsymbol{W}_{obs}, \boldsymbol{W}_{mis})$. We use the generic notation f for the distribution of responses, π for the prior and posterior distributions of related parameters, and L for the likelihood function. Our algorithm consists of a DA imputation step and a posterior sampling step as follows:

(1) *DAI Step* (*data augmentation imputation step*): Sample latent data $\boldsymbol{Y}$ and missing values $\boldsymbol{W}_{mis}$ from $f(\boldsymbol{Y}, \boldsymbol{W}_{mis}|\boldsymbol{\theta}, \boldsymbol{Q}_{obs}, \boldsymbol{W}_{obs})$. To do this, we factor this distribution as

$$\begin{aligned}&f(\boldsymbol{Y}, \boldsymbol{W}_{mis}|\boldsymbol{\theta}, \boldsymbol{Q}_{obs}, \boldsymbol{W}_{obs})\\ =&f(\boldsymbol{W}_{mis}|\boldsymbol{\theta}, \boldsymbol{Y}_{obs}, \boldsymbol{Y}_{mis}, \boldsymbol{W}_{obs})f(\boldsymbol{Y}_{mis}|\boldsymbol{\theta}, \boldsymbol{Y}_{obs}, \boldsymbol{W}_{obs})f(\boldsymbol{Y}_{obs}|\boldsymbol{\theta}, \boldsymbol{Q}_{obs}, \boldsymbol{W}_{obs})\end{aligned} \tag{4.1}$$

(2) *PS step* (*posterior sampling step*): Generate $\boldsymbol{\theta}$ from $f(\boldsymbol{\theta}|\boldsymbol{Y}, \boldsymbol{W})$ using Gibbs sampling.

The DAI-step involves sampling in the order from $[\boldsymbol{Y}_{obs}|\boldsymbol{\theta}, \boldsymbol{Q}_{obs}, \boldsymbol{W}_{obs}]$, $[\boldsymbol{Y}_{mis}|\boldsymbol{\theta}, \boldsymbol{Y}_{obs}, \boldsymbol{W}_{obs}]$, and $[\boldsymbol{W}_{mis}|\boldsymbol{\theta}, \boldsymbol{Y}_{obs}, \boldsymbol{Y}_{mis}, \boldsymbol{W}_{obs}]$. Once we obtain latent data and missing values, we need to sample from full conditional distributions of the components of $\boldsymbol{\theta}$. This can be completed by using the Gibbs sampler. All the full conditional distributions for the Gibbs sampler are derived in the web appendix, Section I (www.amstat.org/publications/ jasa/ supplemental materials).

4.2 Technical Issues

To implement the MCMC algorithm efficiently, there are two technical issues that need to be addressed: 1) Imputing the latent data; 2) Sampling the correlation matrix.

4.2.1 Imputing the Latent Data

One of the challenges with the multivariate probit models is the simulation of latent variables from the TMVN distribution of $\boldsymbol{Y}_{obs}$ given $(\boldsymbol{\theta}, \boldsymbol{Q}_{obs}, \boldsymbol{W}_{obs})$. We propose an algorithm below to sample from this distribution efficiently.

For simplicity, assume that $\boldsymbol{Y}$ is a $T \times 1$ vector and $\boldsymbol{Y} \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})\mathrm{I}_{U_1}$, where $U_1 \in \mathcal{C}^T$ is a truncation region. If we partition $\boldsymbol{Y}$, $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ as $\boldsymbol{Y} = \begin{pmatrix} y_t \\ \boldsymbol{Y}_{(-t)} \end{pmatrix}$, $\boldsymbol{\mu} = \begin{pmatrix} \mu_t \\ \boldsymbol{\mu}_{(-t)} \end{pmatrix}$ and $\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_{tt} & \boldsymbol{\Sigma}_2 \\ \boldsymbol{\Sigma}_2' & \boldsymbol{\Sigma}_{22} \end{pmatrix}$, then we have $y_t | \boldsymbol{Y}_{(-t)} = \boldsymbol{y}_{(-t)} \sim \mathrm{N}(\mu_t^*, \sigma_{tt}^*)\mathrm{I}_{U_{1t}}$ where

$$\mu_t^* = \mu_t + \boldsymbol{\Sigma}_2 \boldsymbol{\Sigma}_{22}^{-1}(\boldsymbol{y}_{(-t)} - \boldsymbol{\mu}_{(-t)}) \quad \text{and} \quad \sigma_{tt}^* = \sigma_{tt} - \boldsymbol{\Sigma}_2 \boldsymbol{\Sigma}_{22}^{-1} \boldsymbol{\Sigma}_2'. \tag{4.2}$$

Here $U_{1t} = \{y_t \in \mathcal{C} : (y_t, \boldsymbol{y}_{(-t)}) \in U_1\}$. Geweke (1991) and Robert (1995) proposed a Gibbs sampling algorithm to sample $\boldsymbol{Y}$. The kernel of the Markov Chain $\boldsymbol{y}^{(k)} = (y_1^{(k)}, \cdots, y_T^{(k)})$ in this algorithm was obtained by successively generating the components of y from their full conditional distributions $\pi(y_t^{(k)} | y_1^{(k)}, \cdots, y_{t-1}^{(k)}, y_{t+1}^{(k-1)}, \cdots, y_T^{(k-1)})$. A disadvantage of this algorithm is that it requires repeatedly computing the T means and variances given in (4.2) and often results in high autocorrelations. The following proposition provides a simple way to sample from the TMVN distribution without the need to compute (4.2) each time and that we expect to provide lower autocorrelations as well.

Proposition 1. *Suppose that $\boldsymbol{Y} \sim TN(\boldsymbol{\mu}, \boldsymbol{\Sigma})I_{U_1}$, where U_1 is the truncation region of $\boldsymbol{Y}$. Decompose $\boldsymbol{\Sigma}$ as $\boldsymbol{P}\boldsymbol{P}'$, where $\boldsymbol{P}$ is a lower triangular matrix. If U_1 is a convex set, then simulating $\boldsymbol{Y}$ from $TN(\boldsymbol{\mu}, \boldsymbol{\Sigma})I_{U_1}$ is equivalent to first sampling $\boldsymbol{Z}$ from*

$TN(\boldsymbol{\nu}, \boldsymbol{I}_{T\times T})I_{U_2}$ and then translating back to $\boldsymbol{Y}$ through $\boldsymbol{Y} = \boldsymbol{PZ}$. Here, $\boldsymbol{\nu} = \boldsymbol{P}^{-1}\boldsymbol{\mu}$ and U_2 is the transformed version of U_1. (See web appendix, Section II for the proof)

This proposition is motivated by the integral calculation method mentioned in Chib and Greenberg (1998, p. 354). The idea behind this proposition is to obtain a more efficient implementation of the Gibbs sampler based on the new set of T conditional distributions of components of $\boldsymbol{Z}$. These distributions are simple in the sense that we do not need to compute means and variances in (4.2) and the univariate truncation intervals U_{2t} can be easily derived. For example, in our model, $\boldsymbol{Y} \sim \mathrm{TN}(\boldsymbol{\mu}, \boldsymbol{R})\mathrm{I}_{U_1}$ with $U_1 = \{\boldsymbol{y} \in \mathcal{C}^T : \boldsymbol{S}_1\boldsymbol{y} \leq \boldsymbol{0}\}$, where $\boldsymbol{S}_1$ is an diagonal matrix with $S_1(t,t) = 1$ if $Q_t = 0$ and -1 if $Q_t = 1$ $(t = 1, \cdots, T)$. By the transformation $\boldsymbol{Z} = \boldsymbol{P}^{-1}\boldsymbol{Y}$, we have $\boldsymbol{Z} \sim \mathrm{TN}(\boldsymbol{\nu}, \boldsymbol{I}_{T\times T})\mathrm{I}_{U_2}$, where $U_2 = \{\boldsymbol{z} \in \mathcal{C}^T : \boldsymbol{S}_2\boldsymbol{z} \leq 0\}$ and $\boldsymbol{S}_2 = \boldsymbol{S}_1\boldsymbol{P}$.

So, at iteration k, $z_t^{(k)}|z_1^{(k)}, \cdots, z_{t-1}^{(k)}, z_{t+1}^{(k-1)}, \cdots, z_T^{(k-1)} \sim \mathrm{N}(\nu_t, 1)\mathrm{I}_{S_t}$, where ν_t is the tth element of $\boldsymbol{\nu}$ and $S_t = \{z_t \in \mathcal{C} : \boldsymbol{S}_2\boldsymbol{z} \leq 0\}$. Let $\boldsymbol{S}_{2(-t)}$ be the matrix $\{s_1, \cdots, s_{t-1}, s_{t+1}, \cdots, s_T\}'$ and $\boldsymbol{z}_{(-t)}$ denote the vector $\{z_1, \cdots, z_{t-1}, z_{t+1}, \cdots, z_T\}'$. Then U_{2t} is given by

$$U_{2t} = \{z_t \in \mathcal{C} : s_t z_t \leq -\boldsymbol{S}_{2(-t)}\boldsymbol{z}_{(-t)}\}. \tag{4.3}$$

The Markov Chain $\boldsymbol{y}^{(k)} = (y_1^{(k)}, \cdots, y_T^{(k)})$ in our implementation of the Gibbs sampler can be obtained by first generating all components of $\boldsymbol{z}$ one by one from $\pi(z_t^{(k)}|z_1^{(k)}, \cdots, z_{t-1}^{(k)}, z_{t+1}^{(k-1)}, \cdots, z_T^{(k-1)})$ $(t = 1, \cdots, T)$, and then translating back to $\boldsymbol{y}^{(k)}$ by $\boldsymbol{y}^{(k)} = \boldsymbol{P}\boldsymbol{z}^{(k)}$.

4.2.2 Sampling the Correlation Matrix

Sampling correlation matrices in MCMC algorithms can be problematic. In addition to the positive definite constraint of covariance matrices, they have diagonal elements fixed at 1. The ideas for data augmentation and parameter expansion, introduced by Liu, Rubin and Wu (1998), Liu and Wu (1999) and van Dyk and Meng (2001) to speed up convergence of algorithms (EM, DA or others), provide a useful tool for this problem. Liu (2007) and Liu and Daniels (2006) developed two-stage parameter expanded reparameterization and Metropolis Hastings (PX-RPMH) algorithms for sampling a correlation matrix $\boldsymbol{R}$ by extending this idea. In these algorithms, the difficulty of simulating $\boldsymbol{R}$ can be overcome by creating an 'expanded' model in which $\boldsymbol{R}$ can be transformed to a less constrained covariance matrix $\boldsymbol{\Psi}$ by borrowing the scale parameters from an expansion parameter matrix. In the following, we derive an PX-RPMH algorithm for sampling the correlation matrix $\boldsymbol{R}_{i,11}$ in the joint models in Section 3. For notational convenience, we denote $\boldsymbol{R}_{i,11}$ as $\boldsymbol{R}_i$ in this Section.

Define $\boldsymbol{\theta}_{(-R_i)}$ to be the parameter vector not including $\boldsymbol{R}_i$, and $\boldsymbol{D}_i$ to be the expansion parameter which is a diagonal matrix we introduce to transform $\boldsymbol{R}_i$ into a less constrained covariance matrix $\boldsymbol{\Psi}_i = \boldsymbol{D}_i\boldsymbol{R}_i\boldsymbol{D}_i$. Consider the following one-to-one mapping from $\{\boldsymbol{Y}_{ij}, \boldsymbol{R}, \boldsymbol{B}_i\}$ to $\{\boldsymbol{Y}^*_{ij}, \boldsymbol{\Psi}_i, \boldsymbol{B}^*_i\}$,

$$\begin{cases} \boldsymbol{Y}_{ij} = \boldsymbol{X}_{1ij}\boldsymbol{\beta} + \boldsymbol{D}_i^{-1}\boldsymbol{Y}^*_{ij} \\ \boldsymbol{R}_i = \boldsymbol{D}_i^{-1}\boldsymbol{\Psi}_i\boldsymbol{D}_i^{-1} \\ \boldsymbol{B}_i = \boldsymbol{B}^*_i\boldsymbol{D}_i \end{cases} \qquad (i = 1, \cdots, m; j = 1, \cdots, n_i), \tag{4.4}$$

where $\sum_{j=1}^{n_i} Y^{*^2}_{ij,t} = 1$ for any $t = 1, \cdots, T$. Given $\boldsymbol{\beta}$, the step that draws $\boldsymbol{Y}_{ij}$ implicitly draws $\boldsymbol{Y}^*_{ij}$ and $\boldsymbol{D}_i$ because $\sum_{j=1}^{n_i}(Y_{ij,t} - \boldsymbol{x}'_{1ij,t}\boldsymbol{\beta})^2 = D^{-1}_{i,tt}\sum_{j=1}^{n_i} Y^{*^2}_{ij,t} = D^{-1}_{i,tt}$, where $D_{i,tt}$ is the tth element of $\boldsymbol{D}_i$ and $\boldsymbol{x}'_{1ij,t}$ is the tth row of $\boldsymbol{X}_{1ij}$. The space for $(\boldsymbol{Y}^*_{ij}, \boldsymbol{\Psi}_i, \boldsymbol{B}^*_i)$ is higher dimensional than that for $(\boldsymbol{Y}_{ij}, \boldsymbol{R}_i, \boldsymbol{B}_i)$ since $\boldsymbol{R}_i$ has fewer parameters than $\boldsymbol{\Psi}_i$. The constraints, $\sum_{j=1}^{n_i} Y^{*^2}_{ij,t} = 1$ for any $t = 1, \cdots, T$, are needed to make the candidate transformation a one-to-one mapping. By specifying the following candidate prior for $\boldsymbol{R}_i$, given by

$$\pi(\boldsymbol{R}_i) \propto |\boldsymbol{R}_i|^{-\frac{a_i}{2}}\mathrm{I}\{r_{i,jk} : r_{i,jk} = 1(j = k), |r_{i,jk}| < 1\ (j \neq k) \text{ and } \boldsymbol{R}_{i,11} \text{ is positive definite}\} \tag{4.5}$$

where a_i is a constant to be determined, we can derive a (parameter expanded) candidate density (PXCD) for $\boldsymbol{\Psi}_i$ based on the following proposition. Note that the candidate prior is introduced solely to derive a candidate density for the Metropolis-Hastings algorithm. It is *not* used for inference.

Proposition 2. *If we choose priors as specified in Section 3.2, then from the likelihood function for the complete data in (3.2), transformation (4.4) and candidate prior (4.5), we obtain*

$$\pi(\boldsymbol{\Psi}_i|\boldsymbol{Y}_i^*, \boldsymbol{B}_i^*, \boldsymbol{\beta}) \propto |\boldsymbol{\Psi}_i|^{-\frac{\nu+T+1}{2}} \exp\Big\{-\frac{1}{2}tr(\boldsymbol{S}\boldsymbol{\Psi}_i^{-1})\Big\}, \tag{4.6}$$

where $\nu_i = n_i - T$, $\boldsymbol{S}_i = \sum_{j=1}^{n_i} \boldsymbol{Y}_{ij}^* \boldsymbol{Y}_{ij}^{*'}$, $\boldsymbol{Y}_i^* = (\boldsymbol{Y}_{i1}^*, \cdots, \boldsymbol{Y}_{in_i}^*)$ *and* $\boldsymbol{Y}_{ij}^* = \boldsymbol{D}_i(\boldsymbol{Y}_{ij} - \boldsymbol{X}_{1ij}\boldsymbol{\beta})$. *That is,* $\boldsymbol{\Psi}_i|\boldsymbol{Y}_i^*, \boldsymbol{B}_i, \boldsymbol{\beta}$ *has an inverse Wishart distribution with degrees of freedom* ν_i *and scale parameter* $\boldsymbol{S}_i$. *(see web appendix, Section II for the proof, www.amstat.org/publications/ jasa/ supplemental materials).*

Proposition 2 gives the parameter expanded candidate density (PXCD) of $\boldsymbol{\Psi}_i$ to use as the proposal density in the Metropolis-Hastings stage. In this stage, we first simulate $\boldsymbol{\Psi}_i$ from (4.6) and then obtain the correlation matrix $\boldsymbol{R}_i$ through the reduction function $P(\boldsymbol{\Psi}_i) = \boldsymbol{D}_i^{-1}\boldsymbol{\Psi}_i\boldsymbol{D}_i^{-1}$. Second, we keep the candidate $\boldsymbol{R}_i$ with probability α_i (the acceptance rate in the M-H algorithm). Sampling $\boldsymbol{R}_i$ based on this algorithm is given in the following theorem.

Theorem 1. *Assume that* $\boldsymbol{\theta}_{(-R_i)}$ *and* $\boldsymbol{R}_i$ *are, a priori, independent, i.e.,* $\pi(\boldsymbol{\theta}_{(-R_i)}, \boldsymbol{R}_i) = \pi(\boldsymbol{\theta}_{(-R_i)})\pi(\boldsymbol{R}_i)$. *If we choose priors as specified in Section 3.2 for* $\boldsymbol{\theta}_{(-R_i)}$ *and* $\boldsymbol{R}_i$*, then under transformation (4.4) and candidate prior (4.5), simulating* $\boldsymbol{R}_i$ *is equivalent to simulating* $\boldsymbol{\Psi}_i$ *first from the inverse Wishart distribution (4.6), and then translating it back to* $\boldsymbol{R}_i$ *through* $\boldsymbol{R}_i = \boldsymbol{D}_i^{-1}\boldsymbol{\Psi}_i\boldsymbol{D}_i^{-1}$ *in (4.4) and accepting the candidate* $\boldsymbol{R}_i$ *using a Metropolis Hastings step with some acceptance rate* α_i*, where* $\alpha_i = min\Big\{1, \exp\Big(\frac{a_i}{2}(\log|\boldsymbol{R}_i| - \log|\boldsymbol{R}_i^{(k)}|)\Big)\Big\}$ *at iteration* $k+1$. *(see web appendix, Section II for the proof, www.amstat.org/publications/ jasa/ supplemental materials).*

Theorem 1 provides a simple way to simulate the correlation matrix in the models proposed in this study.

4.2.3 Efficiency of the New Algorithms

Previous work (Liu, 2007; Liu and Daniels, 2006) has shown that the PX-RPMH algorithm is more efficient than other methods, such as the GG sampler (Ritter and Tanner, 1992) and RW-MH algorithms (Chib and Greenberg, 1998) for sampling a correlation matrix. The performance of the algorithms were compared in detail in these papers and will generalize to our setting, the joint models specified in Section 3.

For sampling from the TMVN as described in Section 5.2.1, we conducted several simulations to compare our method to that the Gibbs sampling technique in Robert (1995). In particular, we evaluated the mixing of Markov chains output from the two algorithms by calculating the lag-n ($n = 1, 2, 3, \ldots$) autocorrelation of each component of $\boldsymbol{Y}$ (the faster the decay of autocorrelation, the faster the mixing). We denote our algorithm and the one from Robert (1995) by LD-A and R-A, respectively. The decay of the autocorrelation for the LD-A algorithm was much faster than in the R-A algorithm. See the web appendix, Section III (www.amstat.org/publications/ jasa/ supplemental materials) for additional details. We provide some summary remarks below.

Remark 1. One computational inefficiency of the R-A for sampling from the TMVN distribution is that we need to repeatedly compute conditional means and variances in (4.2). The LD-A does not require this updating. However, the computational gains from this aspect are minimal.

Remark 2. The mixing of the chain from the R-A grows slower as the the truncation region U_1 gets larger. However, the LD-A has the advantage of fast mixing, regardless of the volume of the U_1. Specifically, when the truncation region approaches infinity (i.e., no truncation), the LD-A provides an iid sample.

Remark 3. The correlation between components of $\boldsymbol{Y}$ has little influence on the mixing of the chain from the LD-A, whereas it affects the performance of the R-A. We saw this when comparing the two algorithms for a variety of choices for $\boldsymbol{\Sigma}$.

5 Model Selection and Goodness of Fit

5.1 Model Comparison

We now consider the problem of comparing alternative models. For the models in Section 3, competing models arise from restrictions on the prior for the association matrix $\boldsymbol{B}_i$, correlation matrices $\boldsymbol{R}_{i,11}$, and/or conditional covariance matrices $\boldsymbol{\Sigma}^*_{i22}$. For example, we might assume that $\boldsymbol{R}_{i,11} = \boldsymbol{R}_{11}$ and $\boldsymbol{\Sigma}^*_{i22} = \boldsymbol{\Sigma}^*_{22}$ when there are not enough data to estimate all the parameters. Although Bayes factors and marginal likelihoods are appealing, these approaches are very difficult to implement in a complex hierarchical model, such as the one proposed in this paper (Liu, 2006). As a result, we derive the deviance information criterion (DIC) (Spiegelhalter *et al.*, 2002) to compare the alternative models. A main reason that we use the DIC is that its computation will be an easy byproduct of the MCMC simulations. However, we note its drawbacks which include lack of invariance to re-parameterization and potential ambiguity in the choice of likelihood.

The DIC is defined as a classical estimate of fit plus a penalty,

$$\mathrm{DIC} = D(\overline{\boldsymbol{\theta}}) + 2p_D = \overline{D(\boldsymbol{\theta})} + p_D,$$

where $\overline{D(\boldsymbol{\theta})}$ is the posterior mean of the deviance and $p_D = \overline{D(\boldsymbol{\theta})} - D(\overline{\boldsymbol{\theta}})$ is the effective number of parameters in the model. Given that we have missing data from subjects dropping out, we use the observed data likelihood as the likelihood to construct the DIC (Celeux et al., 2006; Daniels and Hogan, 2008). Thus, $\overline{D(\boldsymbol{\theta})}$ is defined as

$$\overline{D(\boldsymbol{\theta})} = -2\mathrm{E}_{\boldsymbol{\theta}|\boldsymbol{Q}_{obs},\boldsymbol{W}_{obs}}(\log\, f(\boldsymbol{Q}_{obs}, \boldsymbol{W}_{obs}|\boldsymbol{\theta})) + 2\log\, f(\boldsymbol{Q}_{obs}, \boldsymbol{W}_{obs})$$

and p_D is given by

$$p_D = -2\mathrm{E}_{\boldsymbol{\theta}|\boldsymbol{Q}_{obs},\boldsymbol{W}_{obs}}(\log\, f(\boldsymbol{Q}_{obs}, \boldsymbol{W}_{obs}|\boldsymbol{\theta})) + 2\log\, f(\boldsymbol{Q}_{obs}, \boldsymbol{W}_{obs}|\overline{\boldsymbol{\theta}}),$$

where $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\Sigma})$, $\overline{\boldsymbol{\theta}}$ is the posterior mean. Details on the computation of the DIC for the joint models are given in the web appendix, Section IV (www.amstat.org/publications/ jasa/ supplemental materials).

5.2 Model Checking

The ‘best’ model chosen from models under consideration according to the DIC still may not fit the data well. To check model fit, we will use posterior predictive checks based on a discrepancy function (Gelman et al., 1996). The ’significance’ of these checks is often summarized with the posterior predictive p-value (ppp) (Gelman *etal.*, 1996). However, it is well known that the posterior predictive p-value has the major problem that its distribution under the null tends to concentrate around 0.5 (Robins, van der Vaart and Ventura, 2000). As a result, we check the goodness of fit of the selected joint models based on the calibrated posterior predictive p-value (cppp) (Hjort *et al.*, 2006), defined as

$$\text{cppp}(\boldsymbol{u}^{obs}) = \text{P}\{\text{ppp}(\boldsymbol{u}) \leq \text{ppp}(\boldsymbol{u}^{obs})\}, \tag{5.1}$$

where $\boldsymbol{u} = (\boldsymbol{Q}', \boldsymbol{W}')$, ppp$(\boldsymbol{u}^{obs})$ is the posterior predictive p value derived by Gelman *et al.* (2004) and $\boldsymbol{U}$ has the distribution implied by the prior and the model. Note that $\boldsymbol{u}^{obs}$ corresponds to the observed vector of responses along with the data augmented responses sampled at each iteration (i.e., Q_{mis} and W_{mis}). We discuss the form of the discrepancy function in the example. Hjort *et al.* showed that the distribution of ppp$(\boldsymbol{U})$ is a Uniform(0,1) and the cppp in (5.1) is a proper p value. To calculate ppp$(\boldsymbol{U})$ in (5.1), we need to produce a sample of $\boldsymbol{U}$ first. We can obtain $\boldsymbol{U}$ by first sampling $\boldsymbol{\theta}$ from (3.4) and then drawing $\boldsymbol{U}$ from $f(\boldsymbol{U}|\boldsymbol{\theta})$. The problem arises in sampling $\boldsymbol{\theta}$ from $\pi(\boldsymbol{\theta})$, because in (3.4), we specified flat priors for $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}^*_{i,22}$. We will address this sampling issue next.

To approximately sample $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}^*_{i,22}$ from flat priors, we use the following diffuse priors in place of the improper priors for $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}^*_{i,22}$. As we did for the prior for $\boldsymbol{b}_i$ in Section 3.2.2, we run the independence model with $\boldsymbol{b}_i = 0$ to obtain the posterior mean and covariance of $\boldsymbol{\beta}$ (denoted by $\boldsymbol{\mu}_\beta$ and $\boldsymbol{V}_\beta$) and the posterior mean of $\boldsymbol{\Sigma}^*_{i,22}$ (denoted by $\boldsymbol{\mu}_{\Sigma^*_{i,22}}$). Then, the approximate diffuse prior for $\boldsymbol{\beta}$ can be set as a normal with mean $\boldsymbol{\mu}_\beta$ and covariance matrix $\boldsymbol{V}_\beta$ multiplied by the number of subjects (as with a unit information prior (Kass and Wasserman, 1996)), and the diffuse prior for

$\boldsymbol{\Sigma}^*_{i,22}$ set as an inverse Wishart with degrees of freedom T and scale matrix $\boldsymbol{\mu}_{\Sigma^*_{i,22}}$. Note that to sample $\boldsymbol{R}_{i,11}$ from the uniform prior, we use the algorithm in Joe (2006).

6 Data Analysis

We use the methodology described in Sections 3-5 to analyze the association of longitudinal quit status and weight change in the CTQ II clinical trial described in Section 2. We removed the observations at week 1 and 2 from data since the quit rates in these two weeks were very low (0.00% and 1.10%) due to the design of the study in which subjects were not supposed to try to quit smoking until week 3. Although participants were encouraged to perform makeups for sessions missed, there still existed intermittent missing values in quit status and/or weight gain. In addition, a large number of subjects dropped out before the end of the experiment. In the following, we assume this missingness is ignorable.

For the mean of the two longitudinal responses as a function of covariates, we set $\boldsymbol{\beta}$ in (3.1) to be the vector of means at each time point across treatments; thus, the tth row of the design matrix is a vector of 0's with a 1 in the tth slot. Exploratory analysis suggested setting a_0 in (3.8) to be four; this implies a slower decrease than the raw lag.

6.1 Comparing the Competing Models

We considered several models arising from restrictions on the association matrices $\boldsymbol{B}_i$, the prior for the association matrices $\boldsymbol{B}_i$, and the correlation matrices $\boldsymbol{R}_{i,11}$ and conditional covariance matrices $\boldsymbol{\Sigma}^*_{i,22}$. We fit a total of nine models to the CTQ II trial data. Denote the k-th alternative model by M_k. Table 1 gives the details of all the models considered.

For each of the models in Table 1, we computed the DIC using the methodology derived in Section 5. The DICs for all the models are given in Table 2. From Table 2, we can see that the DIC for model M_5 is smallest and the one for model M_3 is largest. In general, the models using shrinkage priors for association matrices fit better.

6.2 Checking the Goodness of Fit

For the CTQ II data, we defined the discrepancy function to be weekly quit rates or weekly average weight gains. Table 3 gives the cppp for quit rates and average weight gains at each week across treatments for model 5 (M_5). These p values were calculated based on comparison of 2000 pairs of $(\text{ppp}(\boldsymbol{u}), \text{ppp}(\boldsymbol{u}^{obs}))$ with the ppp obtained using $5,000$ iterations after burn-in. From this table, we can see that there are no extreme p-values (< 0.05 or > 0.95). These checks suggest the joint models (M_5) fit the mean structure of the CTQ II clinical trial data well.

6.3 Inference on the Quit Rates and Association

Given the results in Section 6.1 and 6.2, we base inference on model M_5 (cf: Table 1). We ran the MCMC algorithm described in Section 4.1 until convergence (determined by examining trace plots of multiple chains) and based inference on the last 10,000 iterations after burn-in.

Posterior means and 95% credible intervals (CIs) for quit rates across treatments are given in Table 4. The quit rates over time in the exercise treatment were slightly lower than those in the wellness treatment. The 95% credible interval for the difference of quit rates between two treatments at the final week was (-0.014,0.150), marginally significant. This suggests that the exercise treatment does *not* have a positive effect on smoking cessation (more later).

We now turn our attention to the association between smoking cessation and weight gain. The p_i ($i = 1, 2$) defined in (3.9) can be viewed as a summary measure of the overall magnitude of the association between quit status and weight gain for the two treatments. The estimates of p_i are $p_1 = 0.26$ (no exercise) and $p_2 = 0.18$ (exercise); 95% credible interval (CI) for their difference is (0.025,0.134). These results support the hypothesis that exercise weakens the association between quit status and weight gain.

This weakening can also be seen by examining the posterior means of association matrices across treatments as given in Table 5. We have removed from the table those elements of the $\boldsymbol{B}_i$ matrix whose probabilities of the corresponding indicators being equal to 1 is less than 0.1. The weakened associations between smoking cessation and weight gain is obvious by noting the presence of more zeros under the exercise treatment and the larger magnitude of the (standardized) coefficients.

Table 6 shows the posterior means of pairwise correlations with 95% credible intervals; correlations whose 95% credible intervals covered zero were excluded from the table. We can see that smoking cessation and weight gain appear to have a lagged correlation structure, and exercise weakens pairwise correlations. In particular, we point out the two by two blocks in the upper right corners of pairwise correlation matrices under both treatments (they are bolded in the table). For the wellness treatment, the four pairwise correlations between weight gain at the beginning of the study and quit status at weeks 5 and 6 are all negative. This means that people who gain weight early in the trial are more likely to be smoking at the end. The corresponding correlations are essentially zeros (no longer significant) under the exercise arm. In addition, looking at the last row in Table 6 for the wellness arm, the correlations indicate those who quit early in the study are more likely to gain weight by the end (week 6); the corresponding relationship in the exercise arm is weaker.

7 Conclusions and Discussion

We have developed joint models for the association of longitudinal binary and continuous processes and applied them to the analysis of the CTQ II clinical trial to understand the joint evolution of smoking status and weight gain. The results show that moderate-intensity exercise was not successful for smoking cessation but that it did appear to weaken the association between smoking status and weight gain, supporting the hypothesis that exercise has an effect on smoking cessation by weakening the association between quitting smoking and gaining weight.

However, we should be cautious in over-interpreting these results due to the low compliance in the exercise arm. We might expect the low compliance to negatively bias the smoking cessation results (probability of quitting was too low) on the exercise arm. That is, the intention to treat effect (being randomized to the exercise arm) reported here might be expected to be quite different from the causal effect (adhering to the exercise regimen). However, the ability to still see the weakened association between weight gain and smoking on the exercise arm, despite the low compliance, supports this as the mechanism of action for exercise as a therapy for smoking cessation. In future work, we will extend these joint models to estimate causal effects and allow for non-ignorable dropout.

An alternative way to factor model (3.1) is to decompose it into a marginal model for $\boldsymbol{W}_{ij}$ and a conditional probit regression model for $\boldsymbol{Y}_{ij}$, conditioning on $\boldsymbol{W}_{ij}$. This factorization could potentially greatly increase the computational burden in sampling the correlation matrix if the diagonal elements of $\boldsymbol{\Sigma}_{i,11}$ were still fixed at one (due to not being identified). However, this computational problem could be avoided by fixing the diagonal elements of the conditional covariance matrix $\boldsymbol{\Sigma}^{\star}_{i,11} = \boldsymbol{\Sigma}_{i,11} - \boldsymbol{\Sigma}_{i,12}\boldsymbol{\Sigma}^{-1}_{i,22}\boldsymbol{\Sigma}_{i,21}$ to be ones (i.e., make this matrix a correlation matrix). For interpretation of the mean parameters, $\boldsymbol{\beta}$, this computationally simpler approach would require adjusting the $\boldsymbol{\beta}$ components corresponding to the longitudinal binary process with the diagonal elements of the marginal covariance matrix $\boldsymbol{\Sigma}_{i,11}$ which, in this case, would have non-identical diagonal elements.

The general methodology proposed here can be applied to analysis of other data sets where there are two processes, and a question of interest is the association between the two processes. Also the methodology can be directly extended to other longitudinal cases, such as modelling the association between longitudinal ordinal and continuous processes or between two continuous processes (Liu, 2006).

Acknowledgments

This research was supported by NIH grants CA-85295, HL-079457, and CA-77249.

References

Barnard, J., McCulloch, R., and Meng, X.L. (2000). Modeling covariance matrices in terms of standard deviations and correlations with application to shrinkage. *Statistica Sinica*, 10, 1281-1311.

Bartlett, M.S. (1933). On the theory of statistical regression. *Proceedings of the Royal Society of Edinburgh*, 53, 260-283.

Catalano, P.J., and Ryan, L.M. (1992). Bivariate latent variable models for clustered discrete and continuous outcomes. *Journal of the American Statistical Association*, 87, 651-658.

Carlin, B.P., and Chib, S. (1995). Bayesian model choice via Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society, Serial B*, 57, 473-484.

Celeux, G., Forbes, F., Robert, C.P., and Titterington, D.M. (2006). Deviance information criteria for missing data models" (with Discussion), *Bayesian Analysis*, 1, 651-706.

Chen, M.-H., and Shao, Q.-M. (1999). Properties of prior and posterior distributions for multivariate categorical response data models. *Journal of Multivariate Analysis*, 71, 277-296.

Chib, S., and Greenberg, E. (1998). Bayesian analysis of multivariate probit models. *Biometrika*, 85, 347-361.

Chipman, H. (1996). Bayesian variable selection with related predictors. *Canadian Journal of Statistics*, 24, 17-36.

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

Daniels, M.J. (2006). Bayesian modelling of several covariance matrices and some results on propriety of the posterior for linear regression with correlated and/or heterogeneous errors. *Journal of Multivariate Analysis*, 97, 1185-1207.

Daniels, M.J. and Hogan, J.W. (2008) *Missing data in longitudinal studies: Strategies for Bayesian modeling and sensitivity analysis*, Chapman & Hall (CRC Press).

Daniels, M.J., and Kass, R.E. (2001). Shrinkage estimators for covariance matrices. *Biometrics*, 57, 1173-1184.

Daniels, M.J., and Normand S-L (2006). Longitudinal profiling of health care units based on continuous and discrete patient outcomes. *Biostatistics*, 7, 1-15.

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

Dunson, D.B. (2003). Dynamic latent trait models for multidimensional longitudinal data. *Journal of the American Statistical Association*, 98, 555-563.

Fitzmaurice, G.M., and Laird, N.M. (1995). Regression models for a bivariate discrete and continuous outcome with clustering. *Journal of the American Statistical Association*, 90, 845-852.

Gelman, A., Mechelen, I.V., Verbeke, G., Heitjan, D.F., and Meulders, M. (2004). Multiple Imputation for Model Checking: Completed-Data Plots with Missing and Latent Data. *Biometrics*, 61, 74-85.

Gelman, A., Meng, X.-L., and Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies (with discussion). *Statistica Sinica*, 6, 733-807.

George, E.I., and McCulloch, R.E. (1993). Variable selection via Gibbs sampling. *Journal of the American Statistical Association*, 88, 881-889.

George, E.I., and McCulloch, R.E. (1997). Approaches for Bayesian variable selection. *Statistica Sinica*, 7, 339-373

Geweke, J. (1991). Efficient simulation from the multivariate normal and student

t-distributions subject to linear constraints. *Computer Sciences and Statistics*, Proceedings of the 23rd Symposium Interface.

Gueorguieva, R.V., and Agresti, A. (2001). A correlated probit model for joint modelling of clustered binary and continuous responses. *Journal of the American Statistical Association*, 96, 1102-1112.

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

Hjort N.L., Dahl F.A., and Steinbakk G.H. (2006). Post-processing posterior predictive p values. *Journal of the American Statistical Association*, 101, 1157-1174.

Hoeting, J., Raftery, A.E., and Madigan, D. (1996). A method for simultaneous variable selection and outlier identification in linear regression. *Computational Statistics and Data Analysis*, 22, 251-270.

Joe, H. (2006). Generating random correlation matrices based on partial correlations. *Journal of Multivariate Analysis*, 97, 2177-2189.

Kass, R.E., and Wasserman, L. (1996). The selection of prior distributions by formal rules (Corr: 1998V93 p412) *Journal of the American Statistical Association*, 91, 1343-1370.

Liu, X,F. (2006). Bayesian methodology for models with multivariate (longitudinal) outcomes. PhD Dissertation, Department of Statistics, University of Florida.

Liu, X,F. (2007). Parameter expansion for sampling a correlation matrix: an efficient GPX-RPMH algorithm. *Journal of Statistical Computation and Simulation* (in press).

Liu, X.F., and Daniels, M.J. (2006). A New Algorithm for Simulating a Correlation Matrix Based on Parameter Expansion and Re-parameterization. *Journal of Computational and Graphical Statistics*, 16, 897-914.

Liu, C., Rubin, D.B., and Wu, Y.N. (1998). Parameter expansion to accelerate EM: the PX-EM algorithm. *Biometrika*, 85, 755-770.

Liu, J.S., and Wu, Y. (1999). Parameter expansion for data augmentation. *Journal of the American Statistical Association*, 94, 1264-1274.

Marcus, B.H., King, T.K., Albrecht, A.E., Parisi, A.F., and Abrams, D. (1997). Rationale, design, and baseline data for *Commit to Quit*: An exercise efficacy trial for smoking cessation among women. *Preventive Medicine*, 26, 586-597.

Marcus, B.H., Albrecht, A.E., King, T.K., Parisi, A.F., Pinto, B.M., Roberts, M., Niaura, R.S., and Abrams, D.B. (1999). The efficacy of exercise as an aid for smoking cessation in women. *Archives of Internal Medicine*, 36, 479-492.

Marcus, B.H., Lewis, B.A., King, T.K., Albrecht, A.E., Hogan, J., Bock, B., Parisi, A.F., and Abrams, D.B. (2003). Rationale, design, and baseline data for commit to quit II: an evaluation of the efficacy of moderate-intensity physical activity as an aid to smoking cessation in women. *Preventive Medicine*, 36, 479-492.

Marcus, B.H., Lewis, B.A., Hogan, J., King, T.K., Albrecht, A.E., Bock, B., Parisi, A.F., Niaura, R., and Abrams, D.B. (2005). The efficacy of moderate-intensity exercise as an aid for smoking cessation in women: A randomized controlled trial. *Nicotine & Tobacco Research*, 7, 871-880.

Marshall, A.W., and Olkin, I. (1979). *Inequalities–Theory of majorization and applications.* Academic Press: London.

Pourahmadi, M., and Daniels, M.J. (2002). Dynamic conditionally linear mixed models for longitudinal data. *Biometrics*, 58, 225-231.

Regan, M.M., and Catalano, P.J. (1999). Likelihood models for clustered binary and continuous outcomes: application to developmental toxicology. *Biometrics*, 55, 760-768.

Ritter, C., and Tanner, M.A. (1992). Facilitating the Gibbs sampler: The Gibbs stopper and the Griddy-Gibbs sampler. *Journal of the American Statistical Association*, 87, 861-868.

Robert, C.P. (1995). Simulation of truncated normal variables. *Statistics and Computing*, 5, 121-125.

Robins, J.M., Ventura, V., and van der Vaart, A. (2000). Asymptotic distribution of P values in composite null models (Pkg: p1127-1171). *Journal of the American Statistical Association*, 95, 1143-1156.

Roy, J., and Lin, X. (2000). Latent variable models for longitudinal data with multiple continuous outcomes. *Biometrics*, 56, 1047-1054.

Sammel, M.D., Ryan, L.M., and Legler, J.M. (1997). Latent variable models for mixed discrete and continuous outcomes. *Journal of the Royal Statistical Society, Series B*, 59, 667-678.

Smith, M., and Kohn, R. (2002). Parsimonious covariance matrix estimation for longitudinal data. *Journal of the American Statistical Association*, 97, 1141-1153.

Spiegelhalter, D.J., Best, N.G., Carlin, B.P., and van der Linde, A. (2002). Bayesian measures of model complexity and fit (with discussion). *Journal of the Royal Statistical Society, Series B*, 64, 583-639.

Tanner, M.A., and Wong, W.H. (1987). The calculation of posterior distribution by data augmentation (with discussion). *Journal of the American Statistical Association*, 82, 528-550.

van Dyk, D.A., and Meng, X.L. (2001). The art of data augmentation (with discussion). *Journal of Computational and Graphical Statistics*, 10, 1-111.

Wakefield, J.C., and Bennett, J.E. (1996). The Bayesian modelling of covariates for population pharmacokinetic models. *Journal of the American Statistical Association*, 91, 917-927.

Table 1: Models under consideration

Models	Settings			
	$\boldsymbol{B}_i$	Prior for $\boldsymbol{B}_i$	$\boldsymbol{R}_{i,11}$	$\boldsymbol{\Sigma}^*_{i,22}$
M_1	$\boldsymbol{B}_i = 0$	-	Unconstrained	Unconstrained
M_2	$\boldsymbol{B}_i = 0$	-	$\boldsymbol{R}_{i,11} = \boldsymbol{R}_{11}$	$\boldsymbol{\Sigma}^*_{i,22} = \boldsymbol{\Sigma}^*_{22}$
M_3	$\boldsymbol{B}_i = 0$	-	$\boldsymbol{R}_{i,11} = \boldsymbol{I}_{T\times T}$	$\boldsymbol{\Sigma}^*_{i,22} = \mathrm{diag}(\sigma^2_{i1}, \cdots, \sigma^2_{iT})$
M_4	$\boldsymbol{B}_i \neq 0$	Hierarchical Prior	Unconstrained	Unconstrained
$\mathbf{M_5}$	$\mathbf{B_i \neq 0}$	**Hierarchical Prior**	$\mathbf{R_{i,11} = R_{11}}$	$\boldsymbol{\Sigma}^*_{\mathbf{i,22}} = \boldsymbol{\Sigma}^*_{\mathbf{22}}$
M_6	$\boldsymbol{B}_i \neq 0$	Hierarchical Prior	$\boldsymbol{R}_{i,11} = \boldsymbol{I}_{T\times T}$	$\boldsymbol{\Sigma}^*_{i,22} = \mathrm{diag}(\sigma^2_{i1}, \cdots, \sigma^2_{iT})$
M_7	$\boldsymbol{B}_i \neq 0$	Normal Prior*	Unconstrained	Unconstrained
M_8	$\boldsymbol{B}_i \neq 0$	Normal Prior*	$\boldsymbol{R}_{i,11} = \boldsymbol{R}_{11}$	$\boldsymbol{\Sigma}^*_{i,22} = \boldsymbol{\Sigma}^*_{22}$
M_9	$\boldsymbol{B}_i \neq 0$	Normal Prior*	$\boldsymbol{R}_{i,11} = \boldsymbol{I}_{T\times T}$	$\boldsymbol{\Sigma}^*_{i,22} = \mathrm{diag}(\sigma^2_{i1}, \cdots, \sigma^2_{iT})$

$*$ Normal prior is the fractional prior based on (3.8) with $\boldsymbol{\delta}_i = \mathbf{1}$.

Table 2: Estimates of the DIC

	M_1	M_2	M_3	M_4	
DIC	2841.5	2843.0	3775.8	2835.9	
	$\mathbf{M_5}$	M_6	M_7	M_8	M_9
DIC	**2744.6**	3591.3	2834.5	2749.5	3594.1

Table 3: Calibrated posterior predictive p-value for quit rate (QR) and average weight gain (AWG) at each week across treatments for model 5 (M_5)

Treatments	Disc	Weeks					
		1	2	3	4	5	6
Wellness	QR	0.79	0.69	0.12	0.08	0.65	0.27
	AWG	0.23	0.25	0.29	0.17	0.72	0.29
Exercise	QR	0.14	0.26	0.29	0.13	0.18	0.16
	AWG	0.24	0.16	0.26	0.68	0.10	0.18

Table 4: Posterior means and 95% credible intervals (CIs) for quit rates (QR: quit rate; CIL: lower bound; CIU: upper bound)

Items	W					
	1	2	3	4	5	6
QR	**0.40**	**0.50**	**0.58**	**0.53**	**0.53**	**0.46**
CIL	0.34	0.44	0.53	0.47	0.47	0.40
CIU	0.45	0.56	0.66	0.59	0.58	0.52

Items	E					
	1	2	3	4	5	6
QR	**0.45**	**0.37**	**0.44**	**0.42**	**0.46**	**0.38**
CIL	0.40	0.32	0.38	0.36	0.40	0.32
CIU	0.51	0.43	0.49	0.48	0.52	0.43

(W: Wellness treatment; E: Exercise treatment)

Table 5: Posterior means of the association matrices with 95% CIs for each treatment

	Weight	Quit rates					
		1	2	3	4	5	6
W	1	0.39	-0.61	-	2.23	-	-3.08
		(-0.12,0.89)	(-1.19,-0.02)		(1.40,3.06)		(-3.96,-2.20)
	2	-	-	0.40	-	-0.75	-
				(-0.25,1.05)		(-1.48,-0.01)	
	3	-	-	-	-	-	-
	4	-	-	-	-0.32	-	-
					(-1.00,0.37)		
	5	-0.86	-	-	0.95	-	-0.50
		(-1.43,-0.31)			(0.14,1.76)		(-1.23,0.22)
	6	-	-	-	2.13	-1.60	-0.01
					(1.04,3.23)	(-2.68,-0.52)	(-0.80,0.78)
E	1	-	-	-	-	-0.67	-
						(-1.36,0.02)	
	2	-	0.75	-	-	-	-0.90
			(0.18,1.32)				(-1.75,-0.07)
	3	-	-	-	-	-	-
	4	-	-	-	-	-	-
	5	-0.88	-	-	-0.70	-	-
		(-1.50,-0.33)			(-1.33,-0.07)		
	6	-	-1.36	0.83	-	-	-
			(-2.25,-0.49)	(-0.07,1.77)			

(W: Wellness treatment; E: Exercise treatment)

"-" means that the corresponding element is small, defined as $P(\delta = 1 | Q_{obs}, W_{obs}) < 0.1$

Table 6: Posterior means of pairwise correlations with 95 % CIs for each treatment

	Weight	Quit rates					
		1	2	3	4	5	6
W	1	-	-	-0.14	-	**-0.24**	**-0.30**
				(-0.25,-0.02)		(-0.35,-0.13)	(-0.41,-0.19)
	2	-	-	-	-0.12	**-0.16**	**-0.17**
					(-0.24,-0.00)	(-0.28,-0.04)	(-0.29,-0.05)
	3	-	-	-	-	-	-
	4	-	-	-	-	-	-
	5	-	-	-	-	-	-
	6	0.13	0.12	0.15	0.18	-	-
		(0.01,0.24)	(0.01,0.23)	(0.03,0.26)	(0.05,0.30)		
E	1	-	-	-	-	**-**	**-**
	2	-	-	-	-	**-**	**-0.12**
							(-0.24,-0.00)
	3	-	-	-	-	-	-
	4	-	-	-	-	-	-
	5	-	-	-	-	-	-
	6	0.11	-	0.13	-	0.12	0.14
		(0.00,0.22)		(0.03,0.27)		(0.01,0.23)	(0.02,0.25)

(W: Wellness treatment; E: Exercise treatment)

"-" means that the corresponding element is not significant (95% credible interval covers 0)