Published in final edited form as:
J Am Stat Assoc. 2009 June 1; 104(486): 429–438. doi:10.1198/016214508000000904.

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

Xuefeng Liu [Assistant Professor],
Department of Biostatistics and Epidemiology, East Tennessee State University, Johnson City, TN 37614

Michael J. Daniels [Professor; Chair], and
Department of Statistics, University of Florida, Gainesville, FL 32611

Bess Marcus [Professor]
Department of Psychiatry & Human Behavior, Brown University, Providence, RI 02912

Abstract

Joint models for the association of a longitudinal binary and a longitudinal continuous process are proposed for situations in which 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. A Markov chain Monte Carlo (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.

Keywords

Calibrated posterior predictive p-value; Data augmentation; Dependence; Joint models; Markov chain Monte Carlo; Parameter expansion; Stochastic search variable selection

1. INTRODUCTION

In some longitudinal studies, although one time-varying outcome may be of primary interest, several related processes are measured. 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 built 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

mdaniels@stat.ufl.edu.

was to model the association between the two processes. We 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 modeling of mixed outcomes and Bayesian variable selection. For joint modeling, a well-known technique of joint modeling of mixed outcomes is based on introducing a partly observed random variable following a bivariate normal distribution, where one component defines the continuous outcome and the second, latent component defines the binary outcome through the common probit transformation. Authors 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, Ryan, and Legler (1997). In the present work, we 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 and 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 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 that of Carlin and Chib (1995), Chipman (1996), Hoeting, Raftery, and Madigan (1996), and Wakefield and Bennett (1996). George and McCulloch developed stochastic search variable selection (SSVS) to select promising subsets of a set of covariates $X_1, \ldots, X_p$ *for* further consideration in regression models. Smith and Kohn (2002) proposed similar techniques for modeling 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 ideas of Smith and Kohn (2002). The hierarchical specification has the advantage that potential 0's in the association matrix can be identified and estimates of the parameters can be calculated to account for the model uncertainty associated with determining which elements are 0's.

We develop a Markov chain Monte Carlo (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 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 propose better ways to handle these issues.

The article 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 we describe MCMC sampling techniques to estimate the posterior distribution of parameters and address several technical issues regarding efficient implementation of the MCMC algorithm. 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, 2005) was a 4-year randomized controlled clinical trial 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, King, Albrecht, Parisi, and Abrams 1997; Marcus et al. 1999) on the efficacy of vigorous-intensity exercise to aid smoking cessation and weight regulation in women smokers, because moderate-intensity exercise is less arduous and can be performed by healthy individuals without medical supervision. In the CTQ II trial, 217 healthy women aged 18-65 who had regularly smoked 5 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 randomized to one of the two conditions (treatments): a moderate-intensity exercise condition or a contact condition. These two treatments are designated *exercise* and *wellness*. 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 exercise group were required to attend one supervised exercise session per week, 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, which could be performed onsite or at home. Participants in the contact group received lectures, films, and handouts on a variety of health and lifestyle issues. All participants were encouraged to attend makeup sessions if they failed to attend any session during the 8 weeks of treatments. Smoking status was determined through self-report and carbon monoxide testing at each session. In addition, participants were weighed on a weekly basis during the 8 weeks of treatment. 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)—also was 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?

Question 2 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. But our approach also allows us to answer question 1.

3. JOINT MODELS AND PRIORS

3.1 Joint Models

Several authors have developed joint models for the 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. In our setting the time-dependent covariance matrix is modeled as a function of predictors, however, and is of primary interest.

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

$$\mathbf{V}_{ij}=\begin{pmatrix} \boldsymbol{Y}_{ij} \\ \mathbf{W}_{ij} \end{pmatrix}=\mathbf{X}_{ij}\beta+\epsilon_i, \tag{1}$$

where $\mathbf{X}_{ij}$ is the design matrix, $\boldsymbol{\beta}$ is the vector of regression coefficients, and

$$\epsilon_i\sim\mathrm{N}(0,\Sigma_i) \quad \text{with} \quad \Sigma_i=\begin{pmatrix} \Sigma_{i,11} & \Sigma_{i,21} \\ \Sigma_{i,21} & \Sigma_{i,22} \end{pmatrix}.$$

Using the probit formulation for the binary process, we have $Q_{ij,t} = I\{Y_{ij,t} > 0\}$. To estimate the association between $\mathbf{Q}_{ij}$ and $\mathbf{W}_{ij}$, we need models for $\boldsymbol{\Sigma}_{i,12}$ as a function of treatment (and/or other subject specific covariates that might affect this relationship). But both $\boldsymbol{\Sigma}_{i,12}$ and the entire covariance matrix $\boldsymbol{\Sigma}_i$ are difficult to model, because of positive definiteness constraints (Daniels and Kass 2001; Pourahmadi and Daniels 2002) and because it is high-dimensional for each subject. To address this problem, we factor the joint distribution of $\mathbf{Y}_{ij}$ and $\mathbf{W}_{ij}$ into two components: a marginal model for $\mathbf{Y}_{ij}$ and a correlated regression model for $\mathbf{W}_{ij}$ given $\mathbf{Y}_{ij}$, by extending the ideas of Fitzmaurice and Laird (1995) and Gueorguieva and Agresti (2001). Let

$$\mathbf{X}_{ij}=\begin{pmatrix} \mathbf{X}_{1ij} & 0 \\ 0 & \mathbf{X}_{2ij} \end{pmatrix} \quad \text{and} \quad \beta=\begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix};$$

then the new models can be expressed as

$$\boldsymbol{Y}_{ij}=\mathbf{X}_{1ij}\beta_1+\epsilon_{1i} \tag{2}$$

and

$$\mathbf{W}_{ij}=\mathbf{X}_{2ij}\beta_2+\mathbf{B}_i\left(\boldsymbol{Y}_{ij}-\mathbf{X}_{1ij}\beta_1\right)+\epsilon_{2i}. \tag{3}$$

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

It is easy to see that (2) is a correlated probit model and (3) is a standard correlated regression model, conditional on the latent variable $\mathbf{Y}_{ij}$. For identifiability, it is common to restrict $\boldsymbol{\Sigma}_{i,11}$ to be a correlation matrix, $\mathbf{R}_{i,11}$ (Chib and Greenberg 1998); in the rest of this article, we use $\mathbf{R}_{i,11}$ instead of $\boldsymbol{\Sigma}_{i,11}$ as notation for the covariance matrix in (2). The advantage of this factorization is that the components of $\mathbf{B}_i$ in (3) are directly related to the variance and correlation terms in $\boldsymbol{\Sigma}_{i,12}$. In addition, this factorization provides a convenient parameterization for examining the association between $\mathbf{Y}_{ij}$ ($\mathbf{Q}_{ij}$) and $\mathbf{W}_{ij}$, because the components of the $\mathbf{B}_i$ matrix are unconstrained.

Beside being unconstrained, the association matrix $\mathbf{B}_i$ in model (3) can be easily interpreted. The *t*th row of $\mathbf{B}_i$ reflects the association of the continuous process at week *t* with the binary process at all weeks *(t* = 1, . . . , *T)*. In particular, it corresponds to the regression model $w_{ij,t}|\boldsymbol{Y}_{ij}=\mathbf{x}'_{2ij,t}\beta_2+\mathbf{b}'_{i,t}\left(\boldsymbol{Y}_{ij}-\mathbf{X}_{1ij}\beta_1\right)+\epsilon_{2i,t}$ where $\boldsymbol{b}_{i,t}$ = ($b_{i,t1}$, . . . , $b_{i,tT}$)' is the *t*th row of $\mathbf{B}_i$. Because the covariates associated with $\mathbf{b}_{i,t}$, $\mathbf{Y_{ij}}$ - $\mathbf{X}_{1ij}\boldsymbol{\beta}_1$, are centered with variance 1 (recall that the marginal covariance matrix of $\mathbf{Y}_{ij}$ is a correlation matrix), the components of $\mathbf{B}_i$ are *standardized* regression coefficients. This property of the components of $\mathbf{B}_i$ 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 described in Section 3.1. Let $\mathbf{b}_i$ denote the column vector obtained by stringing the rows of $\mathbf{B}_i$ (i = 1, . . . , m); that is, $\mathbf{b}_i$ = ($B_{i,11}$, . . . , $B_{i,1T}$, . . . , $B_{i,TT}$)'. Let $\mathbf{R}_1$ = ($\mathbf{R}'_{1,11}$, . . . , $\mathbf{R}'_{m,11}$)', $\boldsymbol{b}=\left(\boldsymbol{b}'_1,\ldots,\boldsymbol{b}'_{\boldsymbol{m}}\right)'$, and $\Sigma^*_2=\left(\Sigma^{*\prime}_{1,22},\ldots,\Sigma^{*\prime}_{m,22}\right)'$. We write the joint prior in our models as

$$\pi\left(\beta,\mathbf{R}_1,\mathbf{b},\Sigma^*_2\right)=\prod_{i=1}^{m}\pi\left(\beta\right)\pi\left(\mathbf{R}_{i,11}\right)\pi\left(\mathbf{b}_i|\beta,\Sigma^*_{i,22}\right)\pi\left(\Sigma^*_{i,22}\right). \tag{4}$$

This specification implies that $\boldsymbol{\beta}$, $\mathbf{b}_i$, and $\Sigma^*_{i,22}$ are a priori jointly independent of $\mathbf{R}_{i,11}$ and, marginally, $\boldsymbol{\beta}$ and $\Sigma^*_{i,22}$ are a priori independent. Because we have little prior information for $\boldsymbol{\beta}$, $\Sigma^*_{i,22}$ and $\mathbf{R}_{i,11}$, we specify flat priors for $\boldsymbol{\beta}$ and $\Sigma^*_{i,22}$ and a joint uniform prior (derived in Barnard et al. 2000) for $\mathbf{R}_{i,11}$,

$$\pi\left(\beta\right)\propto 1, \tag{5}$$

$$\begin{aligned}\pi\quad&\left(\Sigma^*_{i,22}\right)\propto\mathbf{I}\left\{\Sigma^*_{i,22}\in\left(-\infty,\infty\right)^{T(T+1)/2}\right.\\&\left.\text{and}\quad\Sigma^*_{i,22}\quad\text{is positive definite}\right\},\end{aligned} \tag{6}$$

and

$$\begin{aligned}\pi\quad&\left(\mathbf{R}_{i,11}\right)\propto\mathbf{I}\left\{r_{i,jk}:r_{i,jk}=1\left(j=k\right),|r_{i,jk}|<1\left(j\neq k\right)\right.\\&\text{and}\quad\mathbf{R}_{i,11}\quad\text{is positive definite}\},\end{aligned} \tag{7}$$

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

3.2.1 Prior for the Elements of the Association Matrix—We now provide details on the prior $\pi\left(\mathbf{b}_i|\beta, \Sigma^*_{i,22}\right)$. Because the components of $\mathbf{b}_i$ are the regression coefficients of $\mathbf{W}_{ij}$ on $\mathbf{Y}_{ij}$, we expect many of the components of $\mathbf{b}_i$ to be 0's based on conditional independence (Markov-type) arguments for longitudinal data. Consider, for example, the regression for $w_{ij,t}$. Once we condition on $\{Y_{ij,k} : t - 1 \le k \le t + 1\}$ (current value and lag 1 value forward and backward), we might expect $w_{ij,t}$ to be independent of $\{Y_{ij,k} : k > t + 1, k < t - 1\}$ (values more than lag 1 away). To incorporate these features into the model, we specify a hierarchical prior distribution that essentially allows components of $\mathbf{b}_i$ to be 0's, borrowing ideas from Smith and Kohn (2002). This also will facilitate reducing the number of dependence parameters, which can be quite large.

A key feature of the hierarchical prior is that each component of $\mathbf{b}_i$ (recall that each component is on the same scale because they are standardized regression coefficients) can be exactly 0 with positive probability. To attain this, we introduce the latent indicator vector $\boldsymbol{\delta}_i$ associated with $\mathbf{b}_i$ such that

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

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

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

and covariance matrix

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

where $\mathbf{Z}_{ij} = \mathbf{W}_{ij} - \mathbf{X}_{2ij}\boldsymbol{\beta}_2$, $\mathbf{C}_{ij\delta}$ is obtained by removing from $\mathbf{C}_{ij}$ the columns corresponding to zero elements of $\mathbf{b}_i$, and $\mathbf{C}_{ij}$ is $T \times T^2$ matrix with elements that are functions of $\mathbf{Y}_{ij}$ and the tth row, $\mathbf{C}_{ij,t}=\left(0'_{ij,1}, \ldots, 0'_{ij,(t-1)}, \left(\boldsymbol{Y}_{ij} - \mathbf{X}_{1ij}\beta_1\right)'_t, 0'_{ij,(t+1)}, \ldots, 0'_{ij,T}\right)$.The vector $\boldsymbol{\delta}_{\mathrm{i}}$ is then given the prior

$$\pi\left(\delta_i|p_i\right)=\prod_{t=1}^{T}\prod_{l=1}^{T}\pi\left(\delta_{i,tl|p_i}\right) \tag{8}$$

$$\begin{aligned} = & \prod_{t=1}^{T}\prod_{l=1}^{T}\text{Bernoulli}\left(p_i^{|t-1|^{1/a_{0+1}}}\right), \\ & \pi\left(p_i\right)=\text{Beta}\left(r_i, \lambda_i\right), \end{aligned} \tag{9}$$

where $\boldsymbol{\delta}_i = (\delta_{i,11}, \ldots, \delta_{i,1T}, \ldots, \delta_{i,T1}, \ldots, \delta_{i,TT})'$; $\delta_{i,tl}$ ($i = 1, \ldots, m$; $l, t = 1, \ldots, T$) 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, $\boldsymbol{\lambda}_i$) are the corresponding hyperparameters.

The prior for $\mathbf{b}_i$, given $\boldsymbol{\delta}_i$, is derived based on

$$\pi\left(\mathbf{b}_i|\delta_i,\beta,\Sigma^*_{i,22}\right) \propto f^{1/n_i}\left(\mathbf{W}_i|\boldsymbol{Y}_i,\beta,\mathbf{b}_i,\delta_i,\Sigma^*_{i,22}\right). \quad (10)$$

The rationale for (10) is that the prior provides only 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 (8) can be considered the prior probability that $b_{i,tl}$ will require a nonzero estimate, and $|t - l|^{1/a_0}$ implies that $b_{i,tl}$ is more likely to be 0 as $|t - l|^{1/a_0}$ grows larger. Here a_0 is a tuning parameter that controls the rate of decay of the probability of a nonzero component as a function of lag. Expression (8) implies that the components of $\mathbf{B}_i$ become smaller a priori as they move away from the main diagonal (in the longitudinal setting, become smaller as they move farther away in time). This exponent also can 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 $\mathbf{b}_i$ is that it is a function of $\left(\boldsymbol{Y},\beta,\Sigma^*_{i,22}\right)$, which complicates 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 $\boldsymbol{\mu}\wedge_{bi\delta}$ and covariance $\hat{\boldsymbol{\Sigma}}_{bi\delta}$, defined later. These quantities are computed from an MCMC run with $\mathbf{b}_i = 0$ as $\sum_k \mathrm{E}\left(\mathbf{b}_i|\boldsymbol{Y}^{(k)},\beta^{(k)},\delta,\Sigma^{(k)*}_{22}\right)$ and $\widehat{\Sigma}_{b_{i\delta}}=\sum_k \mathrm{var}\left(\mathbf{b}_i|\boldsymbol{Y}^{(k)},\beta^{(k)},\delta,\Sigma^{(k)*}_{22}\right)$, where k indexes the posterior sample. An alternative would be to just use a mean and variance with the posterior means of $\left(\boldsymbol{Y},\beta,\Sigma^*_{22}\right)$ plugged in.

3.2.3 Propriety of the Joint Posterior Distribution—Given that we have little prior information on $\boldsymbol{\beta}$, $\Sigma^*_{i,22}$ and $\mathbf{R}_{i,11}$, we propose using flat priors for $\boldsymbol{\beta}$, $\Sigma^*_{i,22}$, and $\mathbf{R}_{i,11}$. Because the priors for $\boldsymbol{\beta}$ and $\Sigma^*_{i,22}$ are improper, we need to check that the joint posterior distribution is integrable. In the 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 easily checked conditions are met. The proof of this theorem extends some results of Chen and Shao (1999) for the propriety of posterior distributions for multivariate categorical response models and of Daniels (2006) for the propriety of the posterior for linear regression with cor related 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 $\theta=\left(\beta,\mathbf{R}_1,\mathbf{b},\delta,\mathbf{p},\Sigma^*_2\right)$, where $\delta=\left(\delta'_1,\ldots,\delta'_m\right)'$, $\mathbf{p} = (p_1, \ldots, p_m)'$, and $\mathbf{R}_1$, $\mathbf{b}$, and Σ^*_2 are defined as in (4). To simplify implementation of the algorithm, we include a data augmentation (DA) step (Tanner and Wong 1987) that we use to impute the latent data and the missing values. Let $\mathbf{Q}_{obs}$ be the vector of observed binary outcomes, $\mathbf{Q}_{mis}$ be the vector of missing binary outcomes, $\mathbf{W}_{obs}$ be the vector of observed continuous outcomes, and $\mathbf{W}_{mis}$ be the vector of missing continuous outcomes. Let $\mathbf{Y}_{obs}$ denote the latent vector associated with $\mathbf{Q}_{obs}$ and let $\mathbf{Y}_{mis}$ denote the latent vector related to $\mathbf{Q}_{mis.}$ Define **Y** to be $\left(\boldsymbol{Y}'_{obs}, \boldsymbol{Y}'_{mis}\right)'$ **Q** to be $\left(\mathbf{Q}'_{obs}, \mathbf{Q}'_{mis}\right)'$ and **W** to be $\left(\mathbf{W}'_{obs}, \mathbf{W}'_{mis}\right)'$. We use the generic notation f for the distribution of responses, $\boldsymbol{\pi}$ for the prior and posterior distributions of related parameters, and L for the likelihood function. Our algorithm comprises a DA imputation (DAI) step and a posterior sampling step as follows:

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

$$\begin{aligned}&f\,(\boldsymbol{Y}, \mathbf{W}_{mis}|\theta, \mathbf{Q}_{obs}, \mathbf{W}_{obs})\\&=f\,(\mathbf{W}_{mis}|\theta, \boldsymbol{Y}_{obs}, \boldsymbol{Y}_{mis}, \mathbf{W}_{obs})\,f\,(\boldsymbol{Y}_{mis}|\theta, \boldsymbol{Y}_{obs}, \mathbf{W}_{obs})\\&\quad\times f\,(\boldsymbol{Y}_{obs}|\theta, \mathbf{Q}_{obs}, \mathbf{W}_{obs})\,.\end{aligned} \tag{11}$$

2. PS step (posterior sampling step). Generate $\boldsymbol{\theta}$ from $f(\boldsymbol{\theta}|\mathbf{Y}, \mathbf{W})$ using Gibbs sampling. The DAI step involves sampling in the order from $[\mathbf{Y}_{obs}|\boldsymbol{\theta}, \mathbf{Q}_{obs}, \mathbf{W}_{obs}]$, $[\mathbf{Y}_{mis}|\boldsymbol{\theta}, \mathbf{Y}_{obs}, \mathbf{W}_{obs}]$, and $[\mathbf{W}_{mis}|\boldsymbol{\theta}, \mathbf{Y}_{obs}, \mathbf{Y}_{mis}, \mathbf{W}_{obs}]$. 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 using the Gibbs sampler. All of 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, two technical issues must be addressed: imputing the latent data and 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 $\mathbf{Y}_{obs}$ given $(\boldsymbol{\theta}, \mathbf{Q}_{obs}, \mathbf{W}_{obs})$. We propose an algorithm to sample from this distribution efficiently.

For simplicity, assume that **Y** is a $T \times 1$ vector and $\mathbf{Y} \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})\mathrm{I}_{U1}$, where $U_1 \in C^T$ is a truncation region. If we partition $\mathbf{Y}, \boldsymbol{\mu}$, and $\boldsymbol{\Sigma}$ as

$$\boldsymbol{Y}=\begin{pmatrix} y_t \\ \boldsymbol{Y}_{(-t)} \end{pmatrix}, \mu=\begin{pmatrix} \mu_t \\ \mu_{(-t)} \end{pmatrix},$$

and

$$\Sigma=\begin{pmatrix} \sigma_{tt} & \Sigma_2 \\ \Sigma'_2 & \Sigma_{22} \end{pmatrix},$$

then we have $y_t|\boldsymbol{Y}_{(-t)}=\mathbf{y}_{(-t)}\mathbf{N}\,(\mu_t^*, \sigma_{tt}^*)\,\mathbf{I}_{U_{1t}}$ where

$$\begin{aligned}\mu_t^*&=\mu_t+\Sigma_2\Sigma_{22}^{-1}\,(\mathbf{y}_{(-t)}-\mu_{(-t)})\\\sigma_{tt}^*&=\sigma_{tt}-\Sigma_2\Sigma_{22}^{-1}\Sigma'_2.\end{aligned} \tag{12}$$

Here $U_{1t} = \{y_t \in C : (y_t, \mathbf{y}_{(-t)}) \in U_1\}$. Geweke (1991) and Robert (1995) proposed a Gibbs sampling algorithm for sampling **Y**. The kernel of the Markov chain $\mathbf{y}^{(k)}=\left(y_1^{(k)}, \ldots, y_T^{(k)}\right)$ in this algorithm was obtained by successively generating the components of **y** from their full conditional distributions $\pi\left(y_t^{(k)} | y_1^{(k)}, \ldots, y_{t-1}^{(k)}, \ldots, y_{t+1}^{(k-1)}, \ldots, y_T^{(k-1)}\right)$. A disadvantage of this algorithm is that it requires repeatedly computing the T means and variances given in (12) and often results in high autocorrelations. The following proposition provides a simple way to sample from the TMVN distribution without the need to compute (12) each time; we expect it to provide lower autocorrelations as well.

Proposition 1: Suppose that $\mathbf{Y} \sim \mathrm{TN}(\boldsymbol{\mu}, \boldsymbol{\Sigma})\mathbf{I}_{U_1}$, where U_1 is the truncation region of $\mathbf{Y}$. Decompose $\boldsymbol{\Sigma}$ as $\mathbf{PP}'$, where $\mathbf{P}$ is a lower triangular matrix. If U_1 is a convex set, then simulating $\mathbf{Y}$ from $\mathrm{TN}(\boldsymbol{\mu}, \boldsymbol{\Sigma})\boldsymbol{I}_{U_1}$ is equivalent to first sampling $\mathbf{Z}$ from $\mathrm{TN}(\boldsymbol{\nu}, \mathbf{I}_{\mathrm{T} \times T})\boldsymbol{I}_{U_2}$ and then translating back to $\mathbf{Y}$ through $\mathbf{Y} = \mathbf{PZ}$. Here, $\boldsymbol{\nu} = \mathbf{P}^{-1}\boldsymbol{\mu}$, and U_2 is the transformed version of U_1 (see the Web appendix, Section II for the proof).

This proposition is motivated by the integral calculation method mentioned by 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 $\mathbf{Z}$. These distributions are simple in the sense that we do not need to compute means and variances in (12), and the univariate truncation intervals U_{2t} can be easily derived. For example, in our model, $\mathbf{Y} \sim \mathrm{TN}(\boldsymbol{\mu}, \mathbf{R})\mathbf{I}_{U_1}$ with $U_1 = \{\mathbf{y} \in C^T\!: \mathbf{S}_1\mathbf{y} \le 0\}$, where $\mathbf{S}_1$ is an diagonal matrix with $S_1(t, t) = 1$ if $Q_t = 0$ and -1 if $Q_t = 1$ ($t = 1, \ldots, T$). By the transformation $\mathbf{Z} = \mathbf{P}^{-1}\mathbf{Y}$, we have $\mathbf{Z} \sim \mathrm{TN}(\boldsymbol{\nu}, \mathbf{I}_{T \times T})\mathbf{I}_{U_2}$, where $U_2 = \{\mathbf{z} \in C^T\!:\mathbf{S}_2\mathbf{z} \le 0\}$ and $\mathbf{S}_2 = \mathbf{S}_1\mathbf{P}$.

Thus, at iteration k, $z_t^{(k)} | z_1^{(k)}, \ldots, z_{t-1}^{(k)}, z_{t+1}^{(k-1)}, \ldots, z_t^{(k-1)} \sim N(\nu_t, 1)\,\mathbf{I}_{U_{2t}}$, where ν_t is the tth element of $\boldsymbol{\nu}$ and $\mathrm{U}_{2\mathrm{t}} = \{z_t \in C\!:\mathbf{S}_2\mathbf{z} \le 0\}$. Let $\mathbf{S}_{2(-t)}$ be the matrix $\{\boldsymbol{s}_1, \ldots, \boldsymbol{s}_{t-1}, \boldsymbol{s}_{t+1}, \ldots, \boldsymbol{s}_T\}'$, and let $\mathbf{z}_{(-t)}$ denote the vector $\{z_1, \ldots, z_{t-1}, z_{t+1}, \ldots, z_T\}'$. Then U_{2t} is given by

$$U_{2t} = \{z_t \in C\!:s_t z_t \le -\mathbf{S}_{2(-t)}\mathbf{z}_{(-t)}\}. \tag{13}$$

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

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 addressing 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 $\mathbf{R}$ by extending this idea. In these algorithms, the difficulty of simulating $\mathbf{R}$ can be overcome by creating an “expanded” model in which $\mathbf{R}$ can be transformed to a less constrained covariance matrix $\boldsymbol{\Psi}$ by borrowing the scale parameters from an expansion parameter matrix. In what follows, we derive an PX-RPMH algorithm for sampling the correlation matrix $\mathbf{R}_{i,11}$ in the joint models in Section 3. For notational convenience, we denote $\mathbf{R}_{i,11}$ as $\mathbf{R}_i$ in this section.

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

$$\left\{\begin{array}{l}\boldsymbol{Y}_{ij}=\mathbf{X}_{1ij}\beta+\mathbf{D}_i^{-1}\boldsymbol{Y}_{ij}^{*}\\ \mathbf{R}_i=\mathbf{D}_i^{-1}\Psi_i\mathbf{D}_i^{-1}\\ \mathbf{B}_i=\mathbf{B}_i^{*}D_i\end{array}\right.\quad(i=1,\ldots,m;j=1,\ldots,n_i),\tag{14}$$

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

$$\begin{array}{ll}\pi & (\mathbf{R}_i)\propto|\mathbf{R}_i|^{-a_i/2}\mathbf{I}\left\{r_{i,jk}:r_{i,jk}=1\,(j=k)\,,|r_i,\,jk|<1\,(j\neq k)\right.\\ & \left.\text{and}\quad\mathbf{R}_{i,11}\quad\text{is positive definite}\right\},\end{array}\tag{15}$$

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 (2), transformation (14) and candidate prior (15), we obtain

$$\begin{array}{ll}\pi & \left(\Psi_i|\boldsymbol{Y}_i^{*},\mathbf{B}_i^{*},\beta\right)\propto|\Psi_i|^{-(\nu_i+T+1)/2}\\ & \times\exp\left\{-\frac{1}{2}\mathrm{tr}\left(\mathbf{S}_i\Psi_i^{-1}\right)\right\},\end{array}\tag{16}$$

where $\nu_i = n_i - T$, $\mathbf{S}_i=\Sigma_{j=1}^{n_i}\boldsymbol{Y}_{ij}^{*}\boldsymbol{Y}_{ij}^{*'}$, $\boldsymbol{Y}_i^{*}=\left(\boldsymbol{Y}_{i1}^{*},\ldots,\boldsymbol{Y}_{in_i}^{*}\right)$, and $\boldsymbol{Y}_{ij}^{*}=\mathbf{D}_i\left(\boldsymbol{Y}_{ij}-\mathbf{X}_{1ij}\beta\right)$. That is, $\Psi_i|\boldsymbol{Y}_i^{*},\mathbf{B}_i,\beta$, $\mathbf{B}_i$, $\boldsymbol{\beta}$ has an inverse-Wishart distribution with degrees of freedom ν_i and scale parameter $\mathbf{S}_i$ (see the Web appendix, Section II for the proof).

Proposition 2 gives the 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 (16), and then obtain the correlation matrix $\mathbf{R}_i$ through the reduction function $P(\Psi_i)=\mathbf{D}_i^{-1}\Psi_i\mathbf{D}_i^{-1}$. Second, we keep the candidate $\mathbf{R}_i$ with probability α_i (the acceptance rate in the Metropolis-Hastings algorithm). Sampling $\mathbf{R}_i$ based on this algorithm is given in the following theorem.

Theorem 1: Assume that $\boldsymbol{\theta}_{(-R_i)}$ and $\mathbf{R}_i$ are a priori independent, that is, $\boldsymbol{\pi}(\boldsymbol{\theta}(_{-R_i}))$, $\mathbf{R}_i) = \boldsymbol{\pi}(\boldsymbol{\theta}$ $(_{-R_i})\boldsymbol{\pi}(\mathbf{R}_i)$. If we choose priors as specified in Section 3.2 for $\boldsymbol{\theta}(_{-R_i})$ and $\mathbf{R}_\mathrm{i}$, then, under transformation (14) and candidate prior (15), simulating $\mathbf{R}_i$ is equivalent to simulating $\boldsymbol{\Psi}_i$ first from the inverse-Wishart distribution (16) and then translating it back to $\mathbf{R}_i$ through $\mathbf{R}_i=\mathbf{D}_i^{-1}\Psi_i\mathbf{D}_i^{-1}$ in (14) and accepting the candidate $\mathbf{R}_i$ using a Metropolis-Hastings step with some acceptance rate α_i, where $\alpha_i=\min\left\{1,\exp\left(\frac{a_i}{2}\left(\log|\mathbf{R}_i|-\log|\mathbf{R}_i^{(k)}|\right)\right)\right\}$ at iteration $k + 1$ (see the Web appendix, Section II for the proof).

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. Those authors compared the performance of the algorithms in detail, and the algorithms 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 with the Gibbs sampling technique of Robert (1995). In particular, we evaluated the mixing of Markov chain output from the two algorithms by calculating the lag-n ($n = 1, 2, 3,. . .$) autocorrelation of each component of **Y**. (The faster the decay of autocorrelation, the faster the mixing.) We denote our algorithm by LD-A and that of Robert (1995) by R-A. The decay of the autocorrelation was much faster for the LD-A algorithm than for the R-A algorithm; see the Web appendix, Section III (www.amstat.org/publications/jasa/supplemental_materials) for additional details. We next provide some summary remarks.

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 (12). The LD-A does not require this updating. The computational gains from this aspect are minimal, however.

Remark 2: The mixing of the chain from the R-A grows slower as the truncation region U_1 increases; however, the LD-A has the advantage of fast mixing, regardless of the volume of 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 **Y** has little influence on the mixing of the chain from the LD-A, whereas it affects the performance of the R-A. We see this when comparing the two algorithms for various choices for **Σ**.

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 $\mathbf{B}_i$, correlation matrices $\mathbf{R}_{i,11}$, and/or conditional covariance matrices $\Sigma^*_{i,22}$ For example, we might assume that $\mathbf{R}_{i,11} = \mathbf{R}_{11}$ and $\Sigma^*_{i,22}=\Sigma^*_{22}$ when there are insufficient data to estimate all of 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 article (Liu 2006). As a result, we derive the deviance information criterion (DIC) (Spiegelhalter, Best, Carlin, and van der Linde 2002) to compare the alternative models. A main reason that we use the DIC is that its computation will be a ready byproduct of the MCMC simulations. But we note its drawbacks, which include lack of invariance to reparameterization and potential ambiguity in the choice of likelihood.

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

$$\mathrm{DIC}=D\left(\bar{\theta}\right)+2_{p_D}=\bar{D}(\theta)+p_D,$$

where $\bar{D}(\theta)$ is the posterior mean of the deviance and $p_D=\bar{D}(\theta)-D\left(\bar{\theta}\right)$ 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, Forbes, Robert, and Titterington 2006; Daniels and Hogan 2008). Thus $\bar{D}(\theta)$ is defined as

$$\begin{aligned}\bar{D}(\theta)= \quad & -2E_{\theta|\mathbf{Q}_{obs},\mathbf{W}_{obs}}\left(\log f\left(\mathbf{Q}_{obs},\mathbf{W}_{obs}|\theta\right)\right)\\ & +2 \quad \log f\left(\mathbf{Q}_{obs},\mathbf{W}_{obs}\right)\end{aligned}$$

and p_D is given by

$$\begin{aligned}P_D= \quad & -2E_{\theta|\mathbf{Q}_{obs},\mathbf{W}_{obs}}\left(\log f\left(\mathbf{Q}_{obs},\mathbf{W}_{obs}|\theta\right)\right)\\ & +2\log f\left(\mathbf{Q}_{obs},\mathbf{W}_{obs}|\ \bar{\theta}\right),\end{aligned}$$

where $\boldsymbol{\theta}=(\boldsymbol{\beta},\boldsymbol{\Sigma})$, and $\boldsymbol{\theta}^{-}$ is the posterior mean. Details of 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 use posterior predictive checks based on a discrepancy function (Gelman, Meng, and Stern 1996). The "significance" of these checks often is summarized with the posterior predictive p value (ppp) (Gelman et al. 1996); however, the ppp has the major problem that its distribution under the null tends to concentrate around .5 (Robins, Ventura, and van der Vaart 2000). As a result, we check the goodness of fit of the selected joint models based on the calibrated ppp (cppp) (Hjort, Dahl, and Steinbakk 2006), defined as

$$cppp\left(\mathbf{u}^{obs}\right)=P\left\{ppp\left(\mathbf{u}\right)\le ppp\left(\mathbf{u}^{obs}\right)\right\}, \tag{17}$$

where $\mathbf{u}=(\mathbf{Q}',\mathbf{W}')'$, $ppp(\mathbf{u}^{obs})$ is the ppp derived by Gelman, Mechelen, Verbeke, Heitjan, and Meulders (2004), and $\mathbf{U}$ has the distribution implied by the prior and the model. Note that $\mathbf{u}^{obs}$ corresponds to the observed vector of responses along with the data augmented responses sampled at each iteration (i.e., $\mathbf{Q}_{mis}$ and $\mathbf{W}_{mis}$). We discuss the form of the discrepancy function in the example. Hjort et al. (2006) showed that the distribution of $ppp(\mathbf{U})$ is a Uniform(0,1) and that the cppp in (17) is a proper p value. To calculate $ppp(\mathbf{U})$ in (17), we need to first produce a sample of $\mathbf{U}$. We can obtain $\mathbf{U}$ by first sampling $\boldsymbol{\theta}$ from (4) and then drawing $\mathbf{U}$ from $f(\mathbf{U}|\boldsymbol{\theta})$. The problem arises in sampling $\boldsymbol{\theta}$ from $\boldsymbol{\pi}(\boldsymbol{\theta})$, because in (4), we specified flat priors for $\boldsymbol{\beta}$ and $\Sigma^*_{i,22}$. We address this sampling issue next.

To approximately sample $\boldsymbol{\beta}$ and $\Sigma^*_{i,22}$ from flat priors, we use the following diffuse priors in place of the improper priors for $\boldsymbol{\beta}$ and $\Sigma^*_{i,22}$ As we did for the prior for $\mathbf{b}_\mathrm{i}$ in Section 3.2.2, we run the independence model with $\mathbf{b}_i=0$ to obtain the posterior mean and covariance of $\boldsymbol{\beta}$ (denoted by $\boldsymbol{\mu}_\beta$ and $\mathbf{V}_\beta$) and the posterior mean of $\Sigma^*_{i,22}$ (denoted by $\mu_{\Sigma^*_{i,22}}$). Then the approximate diffuse prior for $\boldsymbol{\beta}$ can be set as a normal with mean $\boldsymbol{\mu}_{\boldsymbol{\beta}}$ and covariance matrix $\mathbf{V}_{\beta}$ multiplied by the number of subjects (as with a unit information prior [Kass and Wasserman 1996]), and the diffuse prior for $\Sigma^*_{i,22}$ can be set as an inverse-Wishart with degrees of freedom T and scale matrix $\mu_{\Sigma^*_{i,22}}$. Note that to sample $\mathbf{R}_{i,11}$ from the uniform prior, we use the algorithm of 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 the data, because the quit rates in these two weeks were very low (0 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 make up missed sessions, 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 what follows, we assume this missingness is ignorable.

For the mean of the two longitudinal responses as a function of covariates, we set $\boldsymbol{\beta}$ in (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 (8) to be 4; 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 matrixes $\mathbf{B}_i$, the prior for the association matrices $\mathbf{B}_i$, the correlation matrices $\mathbf{R}_{i,11}$ and conditional covariance matrices $\Sigma^*_{\mathrm{i},22}$. We fit a total of nine models to the CTQ II trial data. Denote the kth alternative model by M_k. Table 1 gives the details of all of the models considered.

For each model in Table 1, we computed the DIC using the methodology derived in Section 5. The DICs for all of the models are given in Table 2. From Table 2, we can see that the DIC is the smallest for model M_5 and the largest for model M_3. 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 2,000 pairs of [$ppp(\mathbf{u})$, $ppp(\mathbf{u}^{obs})$], with the ppp obtained using 5,000 iterations after burn-in. From this table, we can see that there are no extreme p values (<.05 or >.95). These checks suggest that 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 presented in Sections 6.1 and 6.2, we base inference on model M_5 (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 were slightly lower in the exercise treatment than in the wellness treatment. The 95% CI for the difference of quit rates between two treatments at the final week was (-.014, .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 (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 = .26 (no exercise) and p_2= .18 (exercise); the 95% CI for their difference is (.025, .134). These results support the hypothesis that exercise weakens the association between quit status and weight gain.

This weakening also can be seen by examining the posterior means of association matrixes across treatments, as given in Table 5. We have removed from the table those elements of the $\mathbf{B}_i$ matrix with probabilities of the corresponding indicators being equal to 1 of <.1. The weakened associations between smoking cessation and weight gain is obvious by noting the presence of more 0's 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 0 are excluded. We can see that smoking cessation and weight gain appear to have a lagged correlation structure, and that exercise weakens pairwise correlations. In particular, we point out the 2 × 2 blocks in the upper right corners of pairwise correlation matrixes under both treatments (in bold-type). 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 0's (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 gain insight into 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.

But we should be cautious in overinterpreting these results, because of 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 (randomization to the exercise arm) reported here might be expected to be quite different from the causal effect (adherence to the exercise regimen). But 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 nonignorable dropout.

An alternative way to factor model (1) is to decompose it into a marginal model for $\mathbf{W}_{ij}$ and a conditional probit regression model for $\mathbf{Y}_{ij}$, conditioning on $\mathbf{W}_{ij}$. This factorization could potentially greatly increase the computational burden in sampling the correlation matrix if the diagonal elements of $\mathbf{\Sigma}_{i,11}$ were still fixed at 1 (due to not being identified); however, this computational problem could be avoided by fixing the diagonal elements of the conditional covariance matrix $\Sigma^{*}_{i,11}=\Sigma_{i,11}-\Sigma_{i,12}\Sigma^{-1}_{i,22}\Sigma_{i,21}$ to be l's (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 nonidentical diagonal elements.

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

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

REFERENCES

Barnard J, McCulloch R, Meng XL. Modeling Covariance Matrices in Terms of Standard Deviations and Correlations With Application to Shrinkage. Statistica Sinica 2000;10:1281–1311.

Bartlett MS. On the Theory of Statistical Regression. Proceedings of the Royal Society of Edinburgh 1933;53:260–283.

Carlin BP, Chib S. Bayesian Model Choice via Markov Chain Monte Carlo Methods. Journal of the Royal Statistical Society, Ser. B 1995;57:473–484.

Catalano PJ, Ryan LM. Bivariate Latent Variable Models for Clustered Discrete and Continuous Outcomes. Journal of the American Statistical Association 1992;87:651–658.

Celeux G, Forbes F, Robert CP, Titterington DM. “Deviance Information Criteria for Missing Data Models” (with discussion). Bayesian Analysis 2006;1:651–706.

Chen M-H, Shao Q-M. Properties of Prior and Posterior Distributions for Multivariate Categorical Response Data Models. Journal of Multivariate Analysis 1999;71:277–296.

Chib S, Greenberg E. Bayesian Analysis of Multivariate Probit Models. Biometrika 1998;85:347–361.

Chipman H. Bayesian Variable Selection With Related Predictors. Canadian Journal of Statistics 1996;24:17–36.

Cox DR, Wermuth N. Response Models for Mixed Binary and Quantitative Variables. Biometrika 1992;79:441–461.

Daniels MJ. 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 2006;97:1185–1207.

Daniels, MJ.; Hogan, JW. Missing Data in Longitudinal Studies: Strategies for Bayesian Modeling and Sensitivity Analysis. Chapman & Hall/CRC Press; Boca Ration, FL: 2008.

Daniels MJ, Kass RE. Shrinkage Estimators for Covariance Matrices. Biometrics 2001;57:1173–1184. [PubMed: 11764258]

Daniels MJ, Normand SL. Longitudinal Profiling of Health Care Units Based on Continuous and Discrete Patient Outcomes. Biostatistics 2006;7:1–15. [PubMed: 15917373]

Dunson DB. Bayesian Latent Variable Models for Clustered Mixed Outcomes. Journal of the Royal Statistical Society, Ser. B 2000;62:355–366.

Dunson DB. Dynamic Latent Trait Models for Multidimensional Longitudinal Data. Journal of the American Statistical Association 2003;98:555–563.

Fitzmaurice GM, Laird NM. Regression Models for a Bivariate Discrete and Continuous Outcome With Clustering. Journal of the American Statistical Association 1995;90:845–852.

Gelman A, Mechelen IV, Verbeke G, Heitjan DF, Meulders M. Multiple Imputation for Model Checking: Completed-Data Plots With Missing and Latent Data. Biometrics 2004;61:74–85. [PubMed: 15737080]

Gelman A, Meng X-L, Stern H. "Posterior Predictive Assessment of Model Fitness via Realized Discrepancies" (with discussion). Statistica Sinica 1996;6:733–807.

George EI, McCulloch RE. Variable Selection via Gibbs Sampling. Journal of the American Statistical Association 1993;88:881–889.

George EI, McCulloch RE. Approaches for Bayesian Variable Selection. Statistica Sinica 1997;7:339–373.

Geweke, J. Efficient Simulation From the Multivariate Normal and Student *t*-Distributions Subject to Linear Constraints; Computer Sciences and Statistics, Proceedings of the 23rd Symposium Interface; 1991; p. 571-578.

Gueorguieva RV, Agresti A. A Correlated Probit Model for Joint Modelling of Clustered Binary and Continuous Responses. Journal of the American Statistical Association 2001;96:1102–1112.

Gueorguieva RV, Sanacora G. Joint Analysis of Repeatedly Observed Continuous and Ordinal Measures of Disease Severity. Statistics in Medicine 2006;25:1307–1322. [PubMed: 16217846]

Hjort NL, Dahl FA, Steinbakk GH. Post-Processing Posterior Predictive p Values. Journal of the American Statistical Association 2006;101:1157–1174.

Hoeting J, Raftery AE, Madigan D. A Method for Simultaneous Variable Selection and Outlier Identification in Linear Regression. Computational Statistics and Data Analysis 1996;22:251–270.

Joe H. Generating Random Correlation Matrices Based on Partial Correlations. Journal of Multivariate Analysis 2006;97:2177–2189.

Kass RE, Wasserman L. The Selection of Prior Distributions by Formal Rules. Journal of the American Statistical Association 1996;91:1343–1370. Corr. (1998), 93, 412.

Liu C, Rubin DB, Wu YN. Parameter Expansion to Accelerate EM: The PX-EM Algorithm. Biometrika 1998;85:755–770.

Liu JS, Wu Y. Parameter Expansion for Data Augmentation. Journal of the American Statistical Association 1999;94:1264–1274.

Liu, XF. unpublished doctoral dissertation. University of Florida, Dept. of Statistics; 2006. Bayesian Methodology for Models With Multivariate (Longitudinal) Outcomes.

Liu XF. Parameter Expansion for Sampling a Correlation Matrix: An Efficient GPX-RPMH Algorithm. Journal of Statistical Computation and Simulation 2008;78:1065–1076.

Liu XF, Daniels MJ. A New Algorithm for Simulating a Correlation Matrix Based on Parameter Expansion and Re-Parameterization. Journal of Computational and Graphical Statistics 2006;16:897–914.

Marcus BH, Albrecht AE, King TK, Parisi AF, Pinto BM, Roberts M, Niaura RS, Abrams DB. "The Efficacy of Exercise as an Aid for Smoking Cessation in Women. Archives of Internal Medicine 1999;36:479–492.

Marcus BH, King TK, Albrecht AE, Parisi AF, Abrams D. Rationale, Design, and Baseline Data for Commit to Quit: An Exercise Efficacy Trial for Smoking Cessation Among Women. Preventive Medicine 1997;26:586–597. [PubMed: 9245683]

Marcus BH, Lewis BA, Hogan J, King TK, Albrecht AE, Bock B, Parisi AF, Niaura R, Abrams DB. The Efficacy of Moderate-Intensity Exercise as an Aid for Smoking Cessation in Women: A Randomized Controlled Trial. Nicotine & Tobacco Research 2005;7:871–880. [PubMed: 16298722]

Marcus BH, Lewis BA, King TK, Albrecht AE, Hogan J, Bock B, Parisi AF, Abrams DB. 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 2003;36:479–492. [PubMed: 12649057]

Pourahmadi M, Daniels MJ. Dynamic Conditionally Linear Mixed Models for Longitudinal Data. Biometrics 2002;58:225–231. [PubMed: 11890319]

Regan MM, Catalano PJ. Likelihood Models for Clustered Binary and Continuous Outcomes: Application to Developmental Toxicology. Biometrics 1999;55:760–768. [PubMed: 11315004]

Ritter C, Tanner MA. Facilitating the Gibbs Sampler: The Gibbs Stopper and the Griddy-Gibbs Sampler. Journal of the American Statistical Association 1992;87:861–868.

Robert CP. Simulation of Truncated Normal Variables. Statistics and Computing 1995;5:121–125.

Robins JM, Ventura V, van der Vaart A. Asymptotic Distribution of *P* Values in Composite Null Models. Journal of the American Statistical Association 2000;95:1143–1156.

Roy J, Lin X. Latent Variable Models for Longitudinal Data With Multiple Continuous Outcomes. Biometrics 2000;56:1047–1054. [PubMed: 11129460]

Sammel MD, Ryan LM, Legler JM. Latent Variable Models for Mixed Discrete and Continuous Outcomes. Journal of the Royal Statistical Society, Ser. B 1997;59:667–678.

Smith M, Kohn R. Parsimonious Covariance Matrix Estimation for Longitudinal Data. Journal of the American Statistical Association 2002;97:1141–1153.

Spiegelhalter DJ, Best NG, Carlin BP, van der Linde A. Bayesian Measures of Model Complexity and Fit" (with discussion). Journal of the Royal Statistical Society, Ser. B 2002;64:583–639.

Tanner MA, Wong WH. "The Calculation of Posterior Distribution by Data Augmentation" (with discussion). Journal of the American Statistical Association 1987;82:528–550.

van Dyk DA, Meng XL. The Art of Data Augmentation" (with discussion). Journal of Computational and Graphical Statistics 2001;10:1–111.

Wakefield JC, Bennett JE. The Bayesian Modelling of Covariates for Population Pharmacokinetic Models. Journal of the American Statistical Association 1996;91:917–927.

Table 1

Models under consideration

	Settings			
Models	$\mathbf{B}_i$	**Prior for** $\mathbf{B}_i$	$\mathbf{R}_{i11}$	$\boldsymbol{\Sigma}^{*}_{i,22}$
M_1	$\mathbf{B}_i = 0$	-	Unconstrained	Unconstrained
M_2	$\mathbf{B}_i = 0$	-	$\mathbf{R}_{i,11} = \mathbf{R}_{11}$	$\Sigma^{*}_{i,22} = \Sigma^{*}_{22}$
M_3	$\mathbf{B}_i = 0$	-	$\mathbf{R}_{i,11} = \mathbf{I}_{T \times T}$	$\Sigma^{*}_{i,22} = \mathrm{diag}\left(\sigma^{*}_{i1}, \ldots, \sigma^{2}_{iT}\right)$
M_4	$\mathbf{B}_i \neq 0$	Hierarchical prior	Unconstrained	Unconstrained
M_5	$\mathbf{B}_i \neq 0$	Hierarchical prior	$\mathbf{R}_{i,11} = \mathbf{R}_{11}$	$\Sigma^{*}_{i,22} = \Sigma^{*}_{22}$
M_6	$\mathbf{B}_i \neq 0$	Hierarchical prior	$\mathbf{R}_{i,11} = \mathbf{I}_{T \times T}$	$\Sigma^{*}_{i,22} = \mathrm{diag}\left(\sigma^{*}_{i1}, \ldots, \sigma^{2}_{iT}\right)$
M_7	$\mathbf{B}_i \neq 0$	Normal prior*	Unconstrained	Unconstrained
M_8	$\mathbf{B}_i \neq 0$	Normal prior*	$\mathbf{R}_{i,11} = \mathbf{R}_{11}$	$\Sigma^{*}_{i,22} = \Sigma^{*}_{22}$
M_9	$\mathbf{B}_i \neq 0$	Normal prior*	$\mathbf{R}_{i,11} = \mathbf{I}_{T \times T}$	$\Sigma^{*}_{i,22} = \mathrm{diag}\left(\sigma^{*}_{i1}, \ldots, \sigma^{2}_{iT}\right)$

*The normal prior is the fractional prior based on (8) with $\boldsymbol{\delta}_{\mathrm{i}} = 1$.

Table 2

Estimates of the DIC

	$\boldsymbol{M_1}$	$\boldsymbol{M_2}$	$\boldsymbol{M_3}$	$\boldsymbol{M_4}$	$\boldsymbol{M_5}$	$\boldsymbol{M_6}$	$\boldsymbol{M_7}$	$\boldsymbol{M_8}$	$\boldsymbol{M_9}$
add:	2,841.5	2,843.0	3,775.8	2,835.9	**2,744.6**	3,591.3	2,834.5	2,749.5	3,594.1

NOTE: The best fitting model, **M5**.

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)

		Weeks					
Treatments	**Disc**	**1**	**2**	**3**	**4**	**5**	**6**
Wellness	QR	.79	.69	.12	.08	.65	.27
	AWG	.23	.25	.29	.17	.72	.29
Exercise	QR	.14	.26	.29	.13	.18	.16
	AWG	.24	.16	.26	.68	.10	.18

Table 4

Posterior means and 95% CIs for quit rates

	W						E					
Items	**1**	**2**	**3**	**4**	**5**	**6**	**1**	**2**	**3**	**4**	**5**	**6**
QR	.40	.50	.58	.53	.53	.46	.45	.37	.44	.42	.46	.38
CIL	.34	.44	.53	.47	.47	.40	.40	.32	.38	.36	.40	.32
CIU	.45	.56	.66	.59	.58	.52	.51	.43	.49	.48	.52	.43

NOTE: W, wellness treatment; E, exercise treatment: QR, quit rate; CIL, lower bound; CIU, upper bound.

Table 5

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

		Quit rate					
	Weight	**1**	**2**	**3**	**4**	**5**	**6**
W	1	.39 (-.12, .89)	-.61 (-1.19, .02)	-	2.23 (1.40, 3.06)	-	-3.08 (-3.96, -2.20)
	2	-	-	.40 (-.25, 1.05)	-	-.75 (-1.48, -.01)	-
	3	-	-	-	-	-	-
	4	-	-	-	-.32 (-1.00, .37)	-	-
	5	-.86 (-1.43, -.31)	-	-	.95 (.14, 1.76)	-	-.50 (-1.23, .22)
	6	-	-	-	2.13 (1.04, 3.23)	-1.60 (-2.68, -.52)	-0.1 (-.80, .78)
E	1	-	-	-	-	-.67 (-1.36, .02)	-
	2	-	.75 (.18, 1.32)	-	-	-	-.90 (-1.75, -.07)
	3	-	-	-	-	-	-
	4	-	-	-	-	-	-
	5	-.88 (-1.50, -.33)	-	-	-.70 (1.33, -.07)	-	-
	6	-	-1.36 (-2.25, -.49)	.83 (-.07, 1.77)	-	-	-

NOTE: **W**, wellness treatment; E, exercise treatment “—” means that the corresponding element is small, defined as P (δ = 1| Q_{obs}, W_{obs}) < .1.

Table 6

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

		Quit rates					
	Weight	**1**	**2**	**3**	**4**	**5**	**6**
W	1	-	-	-.14 (-.25, -.02)	-	**-.24** (**-.35, -.13**)	**-.30** (**-.41, -.19**)
	2	-	-	-	-.12 (-.24, 0)	**-.16** (**-.28, -.04**)	**-.17** (**-.29, -0.5**)
	3	-	-	-	-	-	-
	4	-	-	-	-	-	-
	5	-	-	-	-	-	-
	6	.13 (.01, .24)	.12 (.01, .23)	.15 (.03, .26)	.18 (.05, .30)	-	-
E	1	-	-	-	-	-	-
	2	-	-	-	-	-	**-.12** (**-.24, 0**)
	3	-	-	-	-	-	-
	4	-	-	-	-	-	-
	5	-	-	-	-	-	-
	6	.11 (0, .22)	-	.13 (0.03, .27)	-	.12 (.01, .23)	.14 (.02, .25)

NOTE: W, wellness treatment; E, exercise treatment. “—” means that the corresponding element is not significant (95% CI covers 0).