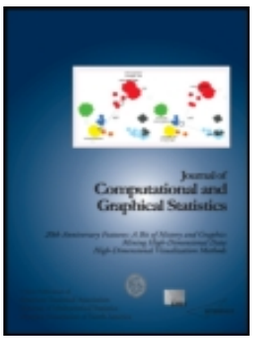

Journal of Computational and Graphical Statistics

ISSN: 1061-8600 (Print) 1537-2715 (Online) Journal homepage: http://www.tandfonline.com/loi/ucgs20

Parameter Expanded Algorithms for Bayesian Latent Variable Modeling of Genetic Pleiotropy Data

Lizhen Xu, Radu V. Craiu, Lei Sun & Andrew D. Paterson

To cite this article: Lizhen Xu, Radu V. Craiu, Lei Sun & Andrew D. Paterson (2016) Parameter Expanded Algorithms for Bayesian Latent Variable Modeling of Genetic Pleiotropy Data, Journal of Computational and Graphical Statistics, 25:2, 405-425, DOI: 10.1080/10618600.2014.988337

To link to this article: http://dx.doi.org/10.1080/10618600.2014.988337

Supplementary materials for this article are available online.
Please go to www.tandfonline.com/r/JCGS

Parameter Expanded Algorithms for Bayesian Latent Variable Modeling of Genetic Pleiotropy Data

Lizhen XU, Radu V. CRAIU, Lei SUN, and Andrew D. PATERSON

Motivated by genetic association studies of pleiotropy, we propose a Bayesian latent variable approach to jointly study multiple outcomes. The models studied here can incorporate both continuous and binary responses, and can account for serial and cluster correlations. We consider Bayesian estimation for the model parameters, and we develop a novel MCMC algorithm that builds upon hierarchical centering and parameter expansion techniques to efficiently sample from the posterior distribution. We evaluate the proposed method via extensive simulations and demonstrate its utility with an application to an association study of various complication outcomes related to Type 1 diabetes. This article has supplementary material online.

Key Words: Bayesian inference; Latent variable; Marginal data augmentation; Markov chain Monte Carlo; Pleiotropy.

1. INTRODUCTION AND MOTIVATION

When the response variable of interest cannot be measured directly we often measure instead a set of surrogate outcomes. The effect of covariates on each observed outcome (also known as manifest variables) can be modeled directly, say via linear or generalized linear models, but the overall effect on the unobserved outcome of interest is difficult to assess. One solution is to use a latent variable (LV) formulation in which the outcome of interest is considered as an unobserved response and can be directly linked to the manifest variables and to the covariates (Bartholomew, Knott, and Moustaki 2011).

Initial applications of LV models focused on reducing the number of manifest variables to a smaller number of latent outcomes. Sammel and Ryan (1996) and Sammel and Ryan (1997) extended the LV methodology to allow covariates to have effects on both the manifest and latent variables. Roy and Lin (2000) discussed a LV approach for longitudinal

Lizhen Xu, Department of Statistical Sciences, University of Toronto (E-mail: *lizhen@utstat.toronto.edu*). Radu V. Craiu, Department of Statistical Sciences, University of Toronto (E-mail: *craiu@utstat.toronto.edu*). Lei Sun, Department of Statistical Sciences and Division of Biostatistics, Dalla Lana School of Public Health, University of Toronto (E-mail: *sun@utstat.toronto.edu*). Andrew D. Paterson, Program in Genetics and Genomic Biology, Hospital for Sick Children, and Dalla Lana School of Public Health, University of Toronto, Toronto (E-mail: *andrew.paterson@utoronto.ca*).

Journal of Computational and Graphical Statistics, Volume 25, Number 2, Pages 405–425
DOI: 10.1080/10618600.2014.988337
Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/r/jcgs.

data with continuous outcomes. Applications of LV modeling appear frequently in a wide spectrum of scientific studies in medicine (Sammel and Ryan 1997), epidemiology (Sanchez et al. 2005), psychology (Engle et al. 1999), and economics (Kuttner 1994), among many others.

Our own interest in latent variable models was motivated by genetics association studies in which a single genetic factor influences multiple continuous or binary phenotypes which are, potentially, different manifestations of the same complex disease. This phenomenon, called pleiotropy, occurs for instance in genetic studies of Type 1 diabetes (henceforth, T1D) where the primary and often conceptual phenotype (e.g., disease severity) may not be directly measured and cannot be characterized by one single phenotype. Instead, subjects may exhibit different levels of renal, retinal, and cardiovascular deterioration. The joint analysis of these surrogate outcomes will increase the statistical efficiency and enhance discovery of genetic risk factors.

An added characteristic of many emerging large-scale genetic studies is the collection of repeated measures over time for clustered units. In genetics the clusters are generally defined by the pedigree/familial structure and are thus assumed known. The longitudinal family studies combine the features of longitudinal studies in independent individuals and studies using single-time-point phenotype measures in families, providing more information about the genetic and environmental factors associated with the traits of interest than cross-sectional studies (Burton et al. 2005). However, joint modeling of multiple phenotypes using longitudinal family data involves nontrivial statistical and computational challenges because of the complex correlations that exist between different phenotypes (the phenotypical correlation), between repeated measures from the same phenotype (the serial correlation) and between individuals within the same family/cluster (the familial correlation).

We consider Bayesian methods that rely on LV models to jointly study multiple correlated outcomes in the presence of serial and cluster correlations. One of the article's contributions is to generalize the work of Roy and Lin (2000) to longitudinal family data that exhibit serial and cluster dependence structures. We discuss the effects of ignoring cluster dependence on the inference for the parameters of interest. We also consider mixed responses that include both binary and continuous phenotypes occurring in unbalanced sampling designs in which the number of observations and the lengths of time intervals between observations vary across subjects.

The Bayesian model we use raises important computational challenges because the posterior distribution is not analytically tractable and, moreover, the standard Markov chain Monte Carlo (MCMC) algorithm used to sample the posterior is inefficient. Another main contribution of the article consists of developing alternative algorithmic designs that improve the sampling efficiency. The MCMC sampler proposed here relies on hierarchical centering and parameter expansion techniques (Gelfand 1995; Liu and Wu 1999; Meng and van Dyk 1999; Hobert and Marchev 2008; Gelman et al. 2008) to improve computational performance.

The rest of the article is organized as follows. Section 2 details the LV model in a general setting. Section 3 presents a Bayesian estimation for the model parameters and a novel MCMC algorithm designed to sample the posterior distribution efficiently. Section 4 shows results from extensive simulation studies, and Section 5 applies the proposed

method to a genetic association study of T1D complications. Section 6 concludes with recommendations and further discussions.

2. STATISTICAL MODEL

We consider here a population of clustered subjects which are measured repeatedly in time. The hierarchical structure of a random sample involves C known clusters, N_c subjects within the cth cluster and K_{ci} repeated measurements for the ith subject from the cth cluster. Let $\mathbf{Y}_{cik} = (y_{cik1}, \ldots, y_{cikJ})^T$ be the $J \times 1$ vector of outcomes (or manifest variables) measured at the kth time point on the ith subject from the cth cluster, for $c = 1, 2, \ldots, C$, $i = 1, 2, \ldots, N_c, k = 1, 2, \ldots, K_{ci}$. Among the J outcomes, $\mathbf{Y}^c_{cik} = (y^c_{cik1}, \ldots, y^c_{cikJ_1})^T$ are continuous and $\mathbf{Y}^b_{cik} = (y^b_{cikJ_1+1}, \ldots, y^b_{cikJ})^T$ are binary.

Let U_{cik} be the LV that represents the underlying overall response which aggregates the partial information brought by each of the J manifest variables and $\mathbf{U}_{ci} = (U_{ci1}, \ldots, U_{ciK_{ci}})^T$ be the vector of the longitudinal LV at times $\mathbf{t}_{ci} = (t_{ci1}, \ldots, t_{ciK_{ci}})^T$.

In the first part of the LV model, a continuous response y^c is linked to the latent trait U via a linear mixed model

$$y^c_{cikj} = \beta_{0j} + \mathbf{W}^T_{cik}\boldsymbol{\beta}_j + \lambda_j U_{cik} + b_{cij} + e_{cikj}, \tag{2.1}$$

where $e_{cikj} \overset{\text{iid}}{\sim} N(0, \sigma^2_j)$, W_{cik} is a p_1-dimensional vector of covariates that have direct effects on the response (also called *direct fixed-effect covariates*) and λ_j is the factor loading that represents the effect of the LV on the jth response. When all λ_j's are equal to 1, model Equation (2.1) is reduced to a mixed effect model. The random component b_{cij} captures the cluster-specific within-subject serial correlations. We assume $b_{cij} \overset{\text{iid}}{\sim} N(0, \tau^2_j)$, and e_{cikj} and b_{cij} are mutually independent for $c = 1, \ldots, C$, $i = 1, \ldots, N_c$, $k = 1, \ldots, K_{ci}$ and $j = 1, \ldots, J$.

If a response is binary, a generalized linear mixed model is assumed,

$$\mu_{cikj} = \beta_{0j} + \mathbf{W}^T_{cik}\boldsymbol{\beta}_j + \lambda_j U_{cik} + b_{cij}, \tag{2.2}$$

with a probit link,

$$E\left[y^b_{cikj}|\mu_{cikj}\right] = \Pr(y^b_{cikj} = 1|\mu_{cikj}) = \Phi(\mu_{cikj}). \tag{2.3}$$

The second part of the LV model specifies the effect of $\mathbf{X}_{cik}$, a set of variables that are of primary interest, on the latent variable $\mathbf{U}$ via a linear mixed model. Elements in X are also called *indirect fixed-effect covariates* because their effects on the response Y are carried out via the effect of the latent variable U on Y in (2.2) or (2.3).

To reflect the correlation implied by the relatedness of subjects within families, we follow the specification of the linear mixed model for family data proposed by Jamsen et al. (2010):

$$\boldsymbol{U}_{ci} = \mathbf{X}_{ci}\boldsymbol{\alpha} + \mathbf{Z}^T_{(g)ci} \otimes \mathbf{1}_{K_{ci}}\boldsymbol{g}_c + \mathbf{Z}^T_{ci} \otimes \mathbf{1}_{K_{ci}}\boldsymbol{a}_c + \boldsymbol{\epsilon}_{ci}, \tag{2.4}$$

where $\otimes$ denotes Kronecker product, $\boldsymbol{\epsilon}_{ci} = (\epsilon_{ci1}, \ldots, \epsilon_{ciK_{ci}})^T$ is the vector of error terms and $\mathbf{X}_{ci} = (X^T_{ci1}, \ldots, X^T_{ciK_{ci}})^T$ is a $K_{ci} \times p_2$ design matrix for the fixed effects α. The

random effect vectors $\mathbf{a}_c = (a_{c1}, \ldots, a_{cN_c})^T$, $\mathbf{g}_c = (g_{c1}, \ldots, g_{cN_c})^T$ account for genetic and environmental factors, respectively, and are independent of the error terms. Their distributions are modeled as $\mathbf{g}_c \sim N_{N_c}(0, \sigma_g^2 \mathrm{I}_{\mathrm{N_c}})$ and $\mathbf{a}_c \sim N_{N_c}(0, \sigma_a^2 \mathrm{I}_{\mathrm{N_c}})$. The indicator $\mathbf{Z}_{(g)c}^T$ is a $N_c \times N_c$ matrix that identifies which related individuals share a common environment, while $\mathbf{Z}_c^T$ is the Cholesky decomposition of the kinship coefficient matrix of the cth family, $\mathbf{K}_c$, that is, $\mathbf{Z}_c^T \mathbf{Z}_c = K_c$. We use $\mathbf{Z}_{(g)ci}$ and $\mathbf{Z}_{ci}$ to denote the ith column of $\mathbf{Z}_{(g)c}$ and $\mathbf{Z}_c$, respectively. For simplicity, we assume that $g_{ci} = g_c$ for all i and c, and $\mathbf{Z}_{(g)c} = \mathbf{1}_{N_c}$, that is, all the related individuals within a family share a common environmental random effect. Thus, we specify the latent variable model as

$$\mathbf{U}_{ci} = \mathbf{X}_{ci}\boldsymbol{\alpha} + g_c \mathbf{1}_{K_{ci}} + \mathbf{Z}_{ci}^T \otimes \mathbf{1}_{K_{ci}} \boldsymbol{a}_c + \boldsymbol{\epsilon}_{ci}. \tag{2.5}$$

When analyzing pleiotropic effects, the covariate of primary interest is the genotype at a genetic marker. In such a setting, pleiotropy is detected if both the α-component corresponding to the effect of genetic marker on the LV and multiple λ's are significant.

To handle the unequally spaced measurements, we assume that the within subject serial correlation of the latent variable U is due to autoregression (Diggle, Liang, and Zeger 1994) and $\epsilon_{ci}(t)$ is a continuous-time Gaussian process with

$$E(\boldsymbol{\epsilon}_{ci}(t)) = 0,\ \mathrm{var}(\boldsymbol{\epsilon}_{ci}(t)) = \sigma_\epsilon^2,\ \mathrm{cov}(\boldsymbol{\epsilon}_{ci}(t_r), \boldsymbol{\epsilon}_{ci}(t_k)) = \sigma_\epsilon^2 \rho^{|t_r - t_k|}, \tag{2.6}$$

where $0 < \rho < 1$ is the correlation coefficent between the within subject error terms that are one time unit apart. That is, we assume that $\epsilon_{ci} \sim N_{K_{ci}}(0, \sigma_\epsilon^2 \mathbf{H}_{ci})$, where $\mathbf{H}_{ci}$ is a $K_{ci} \times K_{ci}$ matrix with the (r, k)th entry equal to $\rho^{|t_r - t_k|}$. Note that if $h \in \mathbf{R}\backslash\{0\}$ is an arbitrary nonzero constant, then one can rewrite Equation (2.1) as

$$y_{cikj}^c = \beta_{0j} + \mathbf{W}_{cik}^T \boldsymbol{\beta}_j + \lambda_j h^{-1} h U_{cik} + b_{cij} + e_{cikj}, \tag{2.7}$$

implying that without any restriction on λ or the variance of ϵ_{cik}, an infinite number of equivalent models can be created. A similar phenomenon appears in the binary response case. To avoid unidentifiability, we assume that (i) $\sigma_\epsilon = 1$; (ii) $\lambda_j \geq 0$; (iii) the set of direct covariates used in (2.2) or (2.3), and the set of indirect covariates used in (2.5) are disjoint, and (iv) the Equation (2.5) does not contain a fixed intercept.

Splitting the available covariates into two disjoint sets that correspond to direct and indirect effects is a delicate step in establishing the LV model. As far as we know, there are no general diagnostic tools available to guide us in this respect. Sammel and Ryan (1996) suggested to include the covariates of primary interest in the indirect effect set and the covariates that are of secondary importance in the direct effect set. In our applications, we want to include as many covariates as possible in the indirect effect set so that we can investigate their association with the LV. A larger indirect set of covariates also implies a more parsimonious model (Khatab and Fahrmeir 2009) which, in the Bayesian context considered here, leads to a reduction of the computational effort required to sample from the posterior distribution. This matter is complicated by the lack of symmetry observed when moving covariates from the indirect to the direct set and vice versa. Specifically, suppose that we define $U_{cik}^* = U_{cik} - \mathbf{X}_{cik}\boldsymbol{\alpha}$ and then we use U_{cik}^* as the LV in (2.1) and (2.2). One can see that switching X from the indirect to the direct set leads to an equivalent model. However, switching covariates from direct to indirect effect set does not lead to

an equivalent model and may produce different conclusions. Simulations performed in Xu (2012) show that model misspecification achieved by transferring a direct effect to the indirect effect set produces a significant increase in the deviance information criterion (DIC, Spiegelhalter et al. 2002) value. Due to these findings, our strategy for the separation of covariates into direct or indirect set is based on scientific reasoning, inferential focus as well as the comparison of DIC differences between the model that includes all the covariates in the direct effect set and the model that moves the investigated covariate into the indirect effect set. Large increases in the DIC value will suggest that it may be more suitable to include the covariate in the direct effect set. An illustration of this principle is presented in Section 5.

2.1 EFFECTS OF IGNORING CLUSTER CORRELATION

A variable measured on units that belong to the same cluster is expected to yield dependent values. In practice, to reduce the analytic complexity and computational burden, one may choose to assume independence and apply existing methods (e.g., Roy and Lin 2000). However, ignoring the cluster dependence structure may result in biased inference for the model parameters. To crystallize the discussion, we assume a simplified case where the responses are all continuous and there are no repeated measures. The LV model becomes

$$y_{cij} = \beta_{0j} + \mathbf{W}_{ci}^T \boldsymbol{\beta}_j + \lambda_j U_{ci} + e_{cij}, \quad \text{and} \quad U_{ci} = \mathbf{X}_{ci}^T \boldsymbol{\alpha} + g_c + \mathbf{Z}_{ci}^T \boldsymbol{a}_c + \epsilon_{ci},$$

where $c = 1, \ldots, C$, $i = 1, \ldots, N_c$ and $j = 1, \ldots, J$ with independent error terms $e_{cij} \sim N(0, \sigma_j^2)$ and $\epsilon_{ci} \sim N(0, 1)$, $\lambda_j > 0$, $g_c \sim N(0, \sigma_g^2)$ and $\boldsymbol{a}_c \sim N(0, \sigma_a^2 \mathbf{I}_{N_c})$.

The variance of the jth response for individual i in family c can be decomposed in terms of the model parameters as

$$\text{var}(y_{cij}) = \sigma_j^2 + \lambda_j^2 \left[\sigma_g^2 + (K_c)_{ii}\sigma_a^2 + 1\right], \tag{2.8}$$

where $(K_c)_{ii}$ is the (i, i)th entry of the kinship coefficient for family c, which is equal to 0.5 for all i and c.

Suppose that we ignore the cluster correlation in the data and propose the model

$$y_{hj} = \beta_{0j} + W_h^T \beta_j + \widetilde{\lambda}_j \widetilde{U}_h + e_{hj}, \quad \text{and} \quad \widetilde{U}_h = X_h^T \widetilde{\alpha} + \epsilon_h,$$

where $h = 1, \ldots, N$ and N is the total sample size. In this case, the variance of the jth response for individual h is decomposed as

$$\text{var}(y_{hj}) = \sigma_j^2 + \widetilde{\lambda}_j^2. \tag{2.9}$$

Comparing (2.8) and (2.9), it is easy to see that $\widetilde{\lambda}_j > \lambda_j$ (since they are constrained to be nonnegative) and $|\widetilde{\alpha}| = \frac{\lambda_j}{\widetilde{\lambda}_j} |\alpha| < |\alpha|$ because $\sigma_g^2 + 0.5\sigma_a^2 + 1 > 1$. Therefore, ignoring cluster correlation can lead to significant underestimation of the absolute value of α, the effect of a covariate on the LV, and overestimation of the value of λ, the effect of the LV on the response in the first part of the LV model. This is consistent with what's reported in the statistical genetics literature in other settings of association studies (e.g., Thornton and McPeek 2010). With longitudinal data we observe similar pattern of bias for the estimations of α and λ, and the simulations in Section 4 show that the bias can be substantial.

3. BAYESIAN MODEL AND COMPUTATION

The data in our model contain the observed continuous and binary outcomes $\mathbf{Y}$, the direct fixed-effect covariates W, the indirect fixed-effect covariates X, and the kinship coefficient related matrix Z. The vector of parameters of interest is $\Theta = (\beta_0, \boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\lambda}, \boldsymbol{\tau}^2, \boldsymbol{\sigma}^2, \sigma_a^2, \sigma_g^2, \rho)^T$ where $\beta_0 = (\beta_{01}, \ldots, \beta_{0J})^T$, $\boldsymbol{\beta} = (\boldsymbol{\beta}_1', \ldots, \boldsymbol{\beta}_J^T)^T$ with $\boldsymbol{\beta}_j = (\beta_{j1}, \ldots, \beta_{jp_1})^T$, $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_{p_2})^T$, $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_J)^T$, $\boldsymbol{\tau}^2 = (\tau_1^2, \ldots, \tau_J^2)^T$ and $\boldsymbol{\sigma}^2 = (\sigma_1^2, \ldots, \sigma_{J_1}^2)^T$. The intractability of the posterior requires the use of MCMC algorithms for statistical inference. Unfortunately, the commonly used priors in probit and linear mixed effects models and a standard sampling scheme lead to a torpidly mixing chain. In the next section, we discuss algorithmic modifications and related prior specifications. The MCMC algorithm follows the data augmentation (DA) principle of Tanner and Wong (1987) and sample alternatively from the posterior distribution given the complete data and from the conditional distribution of the auxiliary data (LV, random effects) given the observed data and the parameter values. We discuss separately the implementation for continuous and binary responses since the modifications to the vanilla DA are different in the two cases.

3.1 PARAMETER EXPANDED DATA AUGMENTATION FOR CONTINUOUS RESPONSES

When conditional conjugate priors are defined for the model parameters, one can use a standard Gibbs (SG) sampler, in which most of the parameters are drawn from their posterior conditional distribution given random effects and all other parameters. For the serial dependence parameter ρ, there is no conjugate prior and the posterior conditional distribution cannot be sampled directly so the chain's updates for ρ are done using a Metropolis-Hastings transition kernel.

Due to high dependence between the components of the Markov chain corresponding to the parameter vector $\boldsymbol{\Theta}$ and the missing and latent data vector $\mathbb{M}$, we observe a very slow mixing of the chain. Some degree of improvement can be obtained by using hierarchical centering (HC) (Gelfand 1995). The HC technique moves the parameters up the hierarchy via model reformulation. Specifically, in Equation (2.1) we shift β_{0j} up the model hierarchy to be the mean of the random effect b and U so that the new random effect and the new latent variable are $b_{cij}^* = \mu_{bj} + b_{cij}$ and $U_{cik}^* = \mu^* + U_{cik}$, respectively, and $\beta_{0j} = \mu_{bj} + \lambda_j * \mu^*$.

Another general strategy devised to overcome the slow convergence problem of Gibbs algorithms is parameter expansion (PX) (Meng and van Dyk 1999; Liu and Wu 1999). The idea behind PX is to introduce auxiliary parameters and/or latent variables in the model and average over all their possible values to produce inference for the original model of interest. As demonstrated by Meng and van Dyk (1999) and Liu and Wu (1999), this apparently circuitous strategy can be highly beneficial, because the larger parameter space allows the Markov chain to move more freely and breaks the dependence between its components. The successful implementation of parameter expansion depends highly on the particular *scheme* being used.

3.1.1 The PX-HC Algorithm for Continuous Outcomes. We introduce auxiliary parameters $\boldsymbol{\xi} = \{\xi_j : 1 \le j \le J\}$, $\mu^* \in \mathbf{R}$ and $\psi \in \mathbf{R}$ and define the following *parameter-*

expanded with hierarchical centering (PX-HC) model:

$$y^c_{cikj} = W^T_{cik}\boldsymbol{\beta}_j + \lambda^*_j U^*_{cik} + \xi_j b^*_{cij} + e_{cikj}, \tag{3.1}$$

with

$$\mathbf{U}^*_{ci} = \mu^* \mathbf{1}_{K_{ci}} + \mathbf{X}_{ci}\boldsymbol{\alpha}^* + g^*_c \mathbf{1}_{K_{ci}} + \mathbf{Z}^T_{ci} \otimes \mathbf{1}_{K_{ci}} \boldsymbol{a}^*_c + \boldsymbol{\epsilon}^*_{ci}, \tag{3.2}$$

where $b^*_{cij} \sim N(\mu^*_{bj}, \tau^{*2}_j)$, $g^*_c \sim N(0, \sigma^{*2}_g)$, $\boldsymbol{a}^*_c \sim N_{N_c}(0, \sigma^{*2}_{\mathrm{a}} \mathrm{I}_{\mathrm{N_c}})$ and $\boldsymbol{\epsilon}^*_{ci} \sim N_{K_{ci}}(0, \psi^2 \mathbf{H}_{ci})$. The parameters in the original and the expanded model are connected through the following linear transformations that depend on the auxiliary parameters:

$$\boldsymbol{\alpha} = \boldsymbol{\alpha}^*/\psi, \quad U_{cik} = \left(U^*_{cik} - \mu^*\right)/\psi, \quad \sigma^2_a = \sigma^{*2}_a/\psi^2, \quad \sigma^2_g = \sigma^{*2}_g/\psi^2,$$
$$\lambda_j = \lambda^*_j \psi, \quad \beta_{j0} = \mu^*_{bj}\xi_j + \lambda^*_j \mu^*, \quad \tau^2_j = \xi^2_j \tau^{*2}_j, \quad \text{for all } 1 \le j \le J.$$

The parameterization of the PX-HC model is mathematically redundant and renders some of the parameters of the extended model unidentifiable. However, this strategy has been shown to improve the computational efficiency of the MCMC algorithms designed to sample from the posterior distribution of the original model (Gelman et al. 2008; Ghosh and Dunson 2009). A significant improvement in computational efficiency is achieved when the estimates of the original parameters are obtained indirectly by the above parameter transformation, as compared to direct estimation.

To maintain the ability to sample from the conditional posterior distribution in the expanded model, conjugate priors must be used also for the auxiliary parameters. These conjugate priors along with the transformations above, lead to specific priors for the parameters defined in the original model. The absolute value of a t-distributed random variable will have a folded-t distribution (Gelman 2006). The priors of the parameters $\boldsymbol{\tau}$ and $\boldsymbol{\lambda}$ belong to this class since $\mathrm{var}(b^*_{cij}) = \tau^{*2}_j$ in our PX-HC model and $\mathrm{var}(b_{cij}) = \tau^2_j$ in the original model which implies $\tau_j = |\xi_j|\tau^*_j$. When the conditional conjugate normal and inverse-Gamma prior are applied to ξ_j and τ^{*2}_j, respectively, the resulting prior for τ_j is the folded-t distribution. Similarly, since $\lambda_j = \lambda^*_j \psi$, a half normal prior assigned to λ^*_j and inverse-Gamma prior to ψ^2 will result in a folded-t prior for λ_j. Other authors have discussed the suitability of folded-t priors in mixed effects and factor analysis models. For instance, Gelman (2006) noted the added flexibility and improved behavior when random effects are small, and Ghosh and Dunson (2009) suggested the use of folded-t priors for the factor loadings in a factor analysis setting.

We consider independent and conjugate priors for the PX-HC model parameters $\boldsymbol{\Theta}^* = (\boldsymbol{\mu}_b{}^*, \mu^*, \boldsymbol{\alpha}^*, \boldsymbol{\beta}, \boldsymbol{\sigma}^2, \sigma^{*2}_g, \sigma^{*2}_a, \psi, \boldsymbol{\tau}^{*2}, \boldsymbol{\lambda}^*, \boldsymbol{\xi}, \rho)'$ as follows:

1. $\boldsymbol{\mu}_b{}^* \overset{\text{iid}}{\sim} N_J(0, 1000\mathbf{I}_J)$, $\mu^* \sim N(0, 1000)$, $\boldsymbol{\alpha}^* \overset{\text{iid}}{\sim} N_{p_2}(\mathbf{0}, 1000 * \mathbf{I}_{p_2})$.
2. $\boldsymbol{\beta}_j \overset{\text{iid}}{\sim} N_{p_1}(\mathbf{0}, 1000 * \mathbf{I}_{p_1})$, $\boldsymbol{\xi} \overset{\text{iid}}{\sim} N_J(0, 1000 * \mathbf{I}_J)$, $\rho \sim \text{Uniform}(0, 1)$.
3. $\sigma^2_j \overset{\text{iid}}{\sim} \text{IG}(0.1, 0.1)$, for $1 \le j \le J$, $\sigma^{*2}_a \sim \text{IG}(0.1, 0.1)$, $\sigma^{*2}_g \sim \text{IG}(0.1, 0.1)$.
4. $\psi^2 \sim \text{IG}(\frac{v_1}{2}, \frac{v_1}{2})$, $\tau^{*2}_j \overset{\text{iid}}{\sim} \text{IG}(\frac{v_2}{2}, \frac{v_2}{2})$, where v_1 and v_2 are the hyperparameters representing the degrees of freedom (df) of the induced folded-t priors for λ_j and τ_j, respectively, for all $1 \le j \le J$. Throughout we set $v_1 = v_2 = 1$.

5. If $TN_+(\mu, \sigma^2)$ denotes the normal distribution with mean μ and variance σ^2 restricted to $(0, \infty)$, then

$$\lambda_j^* \sim TN_+(0, 1),$$

for each $j = 1, \ldots, J$. As $\lambda_j = \psi\lambda_j^*$ there is no loss of generality in assuming a priori that $\text{var}(\lambda_j^*) = 1$ because then $\text{var}(\lambda_j) = \psi^2$.

With these assigned priors, all parameters with the exception of ρ have conditional posterior distributions that can be sampled directly. For ρ we use the reparametrization $\eta = \log(\frac{\rho}{1-\rho})$. The η component of the Markov chain is updated using a random walk Metropolis-Hastings kernel with proposal $N(\eta_{\text{old}}, v^2)$, where v^2 is tuned so that an acceptance rate between 20%–40% is obtained. The supra-index denotes the iteration of the chain, for example, $\boldsymbol{\beta}_j^{(m)}$ are the states at iteration m of those components of the chain that correspond to $\boldsymbol{\beta}_j$. The key steps in updating the algorithm's Markov chain at iteration $m-1$ are:

Step A: Draw $\boldsymbol{\Theta}^{*(m)}$ from $f(\boldsymbol{\Theta}^*|\mathbb{M}^{*(m-1)}, \boldsymbol{y}^c)$.

Step B: Draw all latent variables $\mathbb{M}^{*(m)}$ from $f(\mathbb{M}^*|\boldsymbol{\Theta}^{*(m)}, \boldsymbol{y}^c)$, which involves sampling $\mathbb{M}^* = (U^*, \boldsymbol{b}^*, \boldsymbol{g}^*, \boldsymbol{a}^*)^T$.

After all samples are collected, we transform $\boldsymbol{\Theta}^{*(m)}$ back into $\boldsymbol{\Theta}^{(m)}$, the vector of parameters defined by the original model. The sampling steps involved in Step A and Step B are included in the Appendix, part of the online supplementary material.

3.2 Parameter Expanded DA for Mixed Responses

Suppose that the response vector includes both continuous and binary random variables. Without loss of generality, we assume that the first J_1 outcomes are continuous and the remaining ones are binary. To address concerns involving the MCMC mixing similar to those in the continuous response case, we define the model

$$y_{cikj}^c = W_{cik}^T\boldsymbol{\beta}_j + \lambda_j^* U_{cik}^* + \xi_j b_{cij}^* + e_{cikj}, \quad 1 \le j \le J_1, \tag{3.3}$$

$$p\left(y_{cikj}^b = 1\right) = \Phi\left(W_{cik}^T\boldsymbol{\beta}_j + \lambda_j^* U_{cik}^* + \xi_j b_{cij}^*\right), \quad J_1 + 1 \le j \le J, \tag{3.4}$$

$$\mathbf{U}_{ci}^* = \mu^* \mathbf{1}_{K_{ci}} + \mathbf{X}_{ci}\boldsymbol{\alpha}^* + g_c^* \mathbf{1}_{K_{ci}} + \mathbf{Z}_{ci}^T \otimes \mathbf{1}_{K_{ci}} \boldsymbol{a}_c^* + \boldsymbol{\epsilon}_{ci}^*, \tag{3.5}$$

where $b_{cij}^* \sim N(\mu_{bj}^*, \tau_j^{*2})$, $g_c^* \sim N(0, \sigma_g^{*2})$, $\boldsymbol{a}_c^* \sim N_{N_c}(0, \sigma_a^{*2}\mathbf{I}_{N_c})$, $\boldsymbol{\epsilon}_{ci}^* \sim N_{K_{ci}}(0, \psi^2\mathbf{H}_{ci})$. The prior distributions are the same as in the continuous case for all parameters in (3.3) and (3.5). In addition, for each $j = J_1 + 1, \ldots, J$ we set

$$\boldsymbol{\beta}_j \overset{\text{iid}}{\sim} N_{p_1}(\mathbf{0}, 1000 * \mathbf{I}_{p_1}), \ \lambda_j^* \overset{\text{iid}}{\sim} N(0, 1)1_{\{\lambda_j^* > 0\}}, \ \xi_j^* \overset{\text{iid}}{\sim} N(0, 1000).$$

The form of the probit regression (3.4) leads to conditional posterior distributions that are not available in closed form and thus hinders a direct implementation of the Gibbs sampler. A solution is the DA scheme proposed by Albert and Chib (1993) in which the augmented data $y_{cikj}^{b*} \sim N(\mu_{cikj}, 1)$ is the Gaussian missing variable whose sign is reported by y_{cikj}^b, that is, $y_{cikj}^b = \mathbf{1}_{\{y_{cikj}^{b*} > 0\}}$. The conditional posteriors corresponding to this expanded model (with the exception of ρ) can be directly sampled from. However, in the model defined by Equation (3.3)−Equation (3.5) we have noticed that y_{cikj}^{b*} and some of the model parameters are highly

dependent a posteriori causing a torpid mixing of the chain. Conditional on the auxiliary variables y^{b*}_{cikj}, the posterior conditional distributions are similar to those encountered in the previous section and thus may yield similar bottlenecks. In addition, since the responses y^{b*}_{cikj} are not observed, we need an additional level of parameter expansion. We introduce another working parameter $\boldsymbol{\gamma} = (\gamma_{J_1+1}, \ldots, \gamma_J)^T \in \mathbf{R}^{J-J_1}$, a one-to-one mapping $\widetilde{y}^{b*}_{cikj} = \gamma_j y^{b*}_{cikj}$ and set $\tilde{\beta}_j = \gamma_j \beta_j$, $\tilde{\lambda}^*_j = \gamma_j \lambda^*_j$ and $\tilde{\xi}_j = \gamma_j \xi_j$. A priori, $\gamma_{J_1+1}, \ldots, \gamma_J$ are iid with prior distribution IG(0.1,0.1). In our simulations, the choice of the hyperparameter values for the working parameter priors does not influence the performance of the parameter-expanded samplers.

The resulting *doubly parameter-expanded with hierarchical centering (PX^2−HC)* algorithm corresponds to Scheme 3 of van Dyk and Meng (2001). As suggested by a referee, Scheme 2 of van Dyk and Meng (2001) can be implemented by averaging out some of the added parameters in the model, but in our simulations this modification did not bring a noticeable improvement in efficiency.

Note that when adding the second layer of parameter expansion we do not alter the prior distributions for (β, λ^*, ξ) so that they remain conjugate. Therefore, the conditional priors given γ_j^2 for the transformed parameters $\widetilde{\boldsymbol{\beta}}_j$, $\widetilde{\lambda^*}_j$ and $\widetilde{\xi}_j$ are $N_{p_1}(\mathbf{0}, \gamma_j^2 \Sigma_{\boldsymbol{\beta}})$, $N(0, \gamma_j^2)\mathbf{1}_{\{\widetilde{\lambda}^*_j > 0\}}$, and $N(0, 1000\gamma_j^2)$, respectively. Below we summarize the mth iteration in the Gibbs sampling algorithm, and we provide a complete description in the online Appendix.

Step C: For all parameters and latent variables that are conditionally independent of the binary outcomes (specifically, $\{(\lambda^*_j, \boldsymbol{\beta}_j, \xi_j, b^*_{cij}, \sigma^2_j) : 1 \le j \le J_1\}$, ψ, $\boldsymbol{\alpha}^*$, $\boldsymbol{a}^*$, g^*, μ^*, $\boldsymbol{\mu}_b{}^*$, σ_a^{*2}, σ_g^{*2}, $\boldsymbol{\tau}^{*2}$, ρ) we use the same updating distributions as for the continuous response model.

Step D: For $j = J_1 + 1, \ldots, J$, draw

$$y^{b*(m)}_{cikj} \sim \begin{cases} TN_+(\mu^{*(m)}_{cikj}, 1), & \text{if} \quad y^b_{cikj} = 1 \\ TN_-(\mu^{*(m)}_{cikj}, 1), & \text{if} \quad y^b_{cikj} = 0 \end{cases},$$

where $\mu^{*(m)}_{citj} = W^T_{cik}\beta^{(m-1)}_j + \lambda^{*(m-1)}_j U^{*(m-1)}_{cik} + \xi^{(m-1)}_j b^{*(m-1)}_{cij}$. Transform y^{b*}_{cikj} to $\widetilde{y}^{b*}_{cikj}$ via $\widetilde{y}^{b*}_{cikj} = \gamma_j y^{b*}_{cikj}$.

The order of updating $(\beta^{(m)}_j, \lambda^{*(m)}_j, \xi^{(m)}_j, \gamma^{2(m)}_j)$ involves sampling first $\gamma^{2(m)}_j$ and then $(\widetilde{\beta}^{(m)}_j, \widetilde{\lambda}^{*(m)}_j, \widetilde{\xi}^{(m)}_j)$ from their conditional densities.

We set $\beta^{(m)}_j = \widetilde{\beta}^{(m)}_j / \gamma^{(m)}_j$, $\lambda^{*(m)}_j = \widetilde{\lambda}^{*(m)}_j / \gamma^{(m)}_j$, and $\xi^{(m)}_j = \widetilde{\xi}^{(m)}_j / \gamma^{(m)}_j$ and then we continue updating all the other parameters and latent variables as detailed in the online Appendix.

After all samples are collected we transform the vectors of parameters for the PX^2-HC model back to the vector of parameters used in the original model.

4. SIMULATION STUDIES

In our simulations we set out to explore the performance of our methods in the general settings in which we have clustered data measured longitudinally at unequally spaced time points. We consider 100 families (clusters) with similar pedigree structure as the one specified in Jiang and McPeek (2014), but having the number of children in the

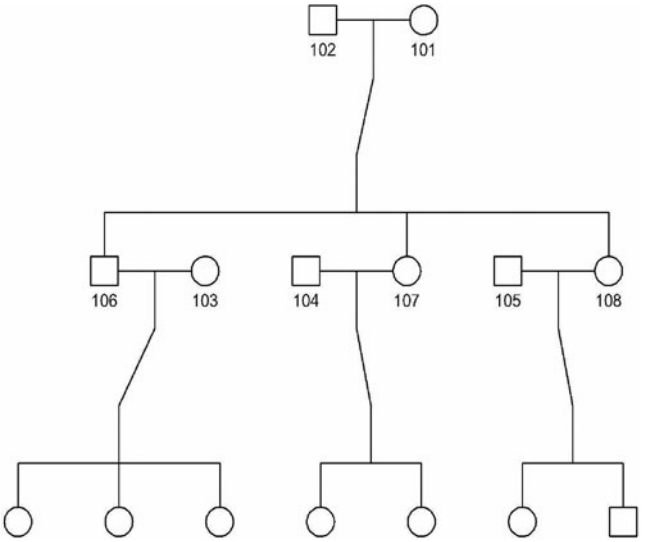

Figure 1. An example of pedigree structure used in simulations.

third generation varying from one to five with probability {20%, 40%, 30%, 7%, 3%}, respectively. An example of the pedigree structure is shown in Figure 1. For each individual, we assume that the probability of being observed longitudinally {1, 2, 3, 4} times is {10%, 30%, 30%, 30%}. The time of first measure is set as {0, 1, 1.5, 2} with probability {50%, 20%, 20%, 10%}, respectively. The length of time between two consecutive measures is {1, 2, 3, 3.5} with probability {50%, 20%, 20%, 10%}, respectively, resulting in an unbalanced design. The serial dependence is modeled via Equation (2.6). The code is included in the online supplementary material.

We have run the three algorithms considered—SG, PX-HC, and PX^2-HC—on two simulation models. In both studies, we assume that there is only one direct effect covariate following a $N(0, 1)$ distribution. We also assume that there are two indirect effect covariates in the model, with the first one following a $N(0, 1)$ distribution and the second one being the genotype of the SNP under study. The individual genotypes are generated using an additive genetic model with the minor allele frequency (MAF) set to 0.3. The genotypes of the founders from the 100 families follow Hardy Weinberg equilibrium (HWE) and alleles pass to next generation according to Mendel's law of segregation. The parameter values for each model were chosen as follows:

M1: We consider $J = 3$ continuous response variables and set $\boldsymbol{\beta}_0 = (5, 5, 5)$, $\beta_{11} = \beta_{12} = \beta_{13} = 1$, $\alpha_1 = -1$, $\alpha_2 = 1$, $\boldsymbol{\lambda} = (5, 5, 5)$, $\boldsymbol{\tau}^2 = (0.3, 0.3, 0.3)$, $\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = 1$, $\sigma_a^2 = 0.3$, $\sigma_g^2 = 0.3$, and $\rho = 0.3$.

M2: We consider $J = 4$ and we simulate y_1, y_2 as continuous and y_3, y_4 as binary responses. We set $\boldsymbol{\beta}_0 = (1, 1, 1, 1)$, $\beta_{1j} = 1$ for all $j = 1, \ldots, 4$, $\alpha_1 = -1$, $\alpha_2 = 1$, $\boldsymbol{\lambda} = (2, 3, 1, 1)$, $\boldsymbol{\tau}^2 = (0.6, 0.6, 0.6, 0.6)$, $\sigma_1^2 = \sigma_2^2 = 1$, $\sigma_a^2 = 1$, $\sigma_g^2 = 1$, $\rho = 0.3$.

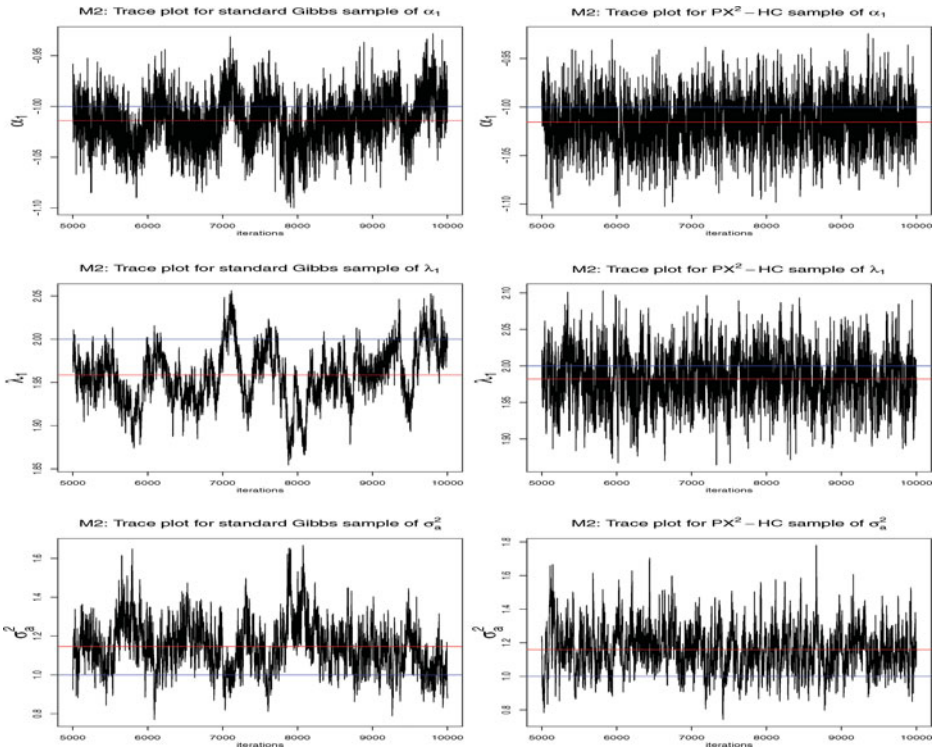

Figure 2. Comparison of trace plots for simulations under model M2 using SG and PX^2-HC scheme. The blue line marks the true value of the parameter, and the red line represents the posterior mean. Left side from top to bottom: trace plots for α_1, λ_1, and σ_a^2 using SG. Right side from top to bottom: trace plots for α_1, λ_1, and σ_a^2 using PX^2 – HC.

4.1 GRAPHICAL EVIDENCE OF IMPROVEMENT

The improved mixing of the Markov chains corresponding to the modified algorithms can be noticed graphically from trace plots, autocorrelation plots and convergence diagnostic plots. In Figure 2 we compare the trace plots for α_1, λ_1, and σ_a^2 using draws from the posterior under M2 obtained via the SG and PX^2-HC algorithms. Additional trace plots for models M1 are included in the online Appendix. We have consistently observed that PX^2-HC is more efficient than SG. The improvements brought by PX^2-HC are more significant for those components of the SG chain that exhibit sluggish mixing and do not slow down the components that are mixing well. The change brought by PX^2-HC is clearly represented visually by the autocorrelation functions (ACF) which present the strength of dependence between successive Monte Carlo draws. This dependence plays an important role when assessing the Monte Carlo error of the samplers (Geyer 1992; Flegal, Haran, and Jones 2008). Figure 3 shows the reduction in autocorrelation for two factor loadings (λ_1 corresponds to a continuous response and λ_3 to a binary one). Each plot contains 100 ACF curves obtained from 100 independent replicates of simulated data under scenario M2. The green curves represent the performance of PX^2-HC with average ACF plotted in blue, while the SG counterpart curves are plotted in purple with their average represented

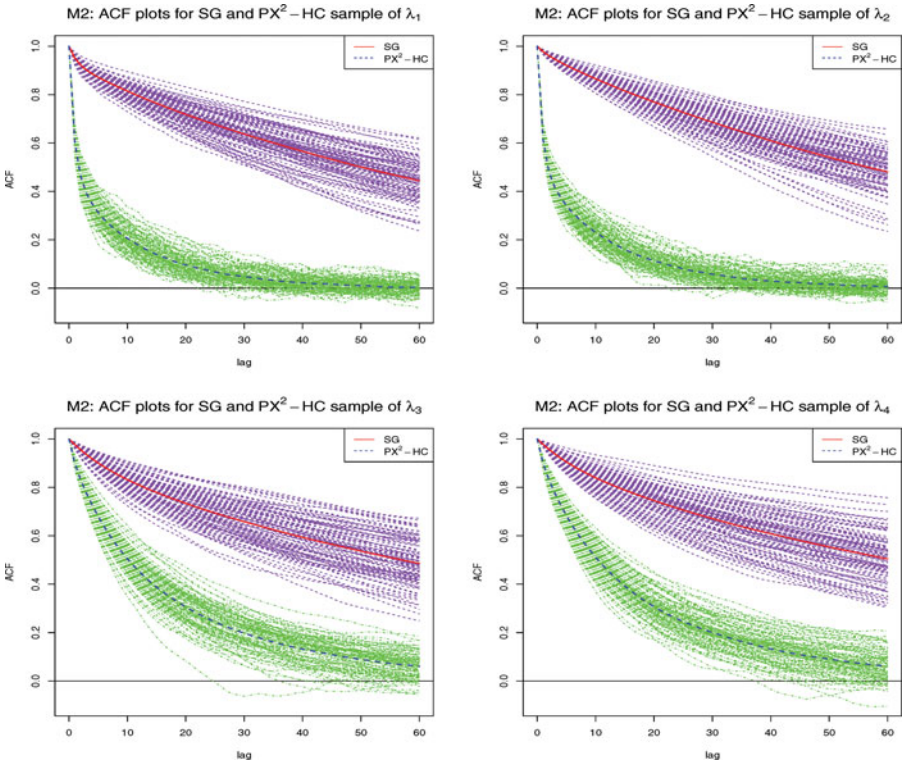

Figure 3. Comparison of ACF plots for the four loading factors λ_j, $j = 1, \ldots, 4$ for model M2. Red line shows the average ACF curve for SG computed from 100 replicated curves which are shown in purple. The blue line shows the average ACF curve for PX^2 − HC computed from 100 replicated curves which are shown in green.

by the red curve. The improved mixing influences also the convergence diagnostic plots. We have followed the general principles of Gelman and Rubin (1992) and ran in parallel 5 chains that were started at random points drawn from a highly spread out distribution. The diagnostic plots for models M1 and M2, included in the Appendix, show the evolution of R^2 for factor loadings λ_1 and λ_3 based on draws obtained with the SG and PX-HC under scenario M1 and by SG and PX^2-HC under M2 (results for λ_2 and λ_4 are characteristically similar). It can be noticed that the modified algorithms have the R^2 approaching 1 earlier in the simulation process.

4.2 Efficiency Comparison

In Tables 1–2 we report the gain in efficiency for α, λ, and the random effect variances σ_a^2 and σ_g^2, when using PX-HC or PX^2-HC versus SG, in terms of root mean squared error (RMSE) and effective sample size (ESS). For comparing algorithms A_1 and A_2 we compute, for each parameter in the table,

$$\Delta_{\text{RMSE}}(A_1, A_2) = 100 \times \left(\frac{\text{RMSE}_{A_1} - \text{RMSE}_{A_2}}{\text{RMSE}_{A_1}} \right) \tag{4.1}$$

Table 1. Simulation results under model M1: parameter estimation comparison using SG and PX-HC

Parameters		Value	SG		PX-HC			
			Est.	RMSE	Est	RMSE	Δ_{RMSE}	Δ_{ESS}
α	α_1	−1.0	−1.003	0.024	−1.000	0.023	<5	1501
	α_2	1.0	1.006	0.044	1.003	0.043	<5	392
λ	λ_1	5.0	4.991	0.086	5.007	0.079	8	1942
	λ_2	5.0	4.994	0.084	5.009	0.077	6	1968
	λ_3	5.0	4.991	0.086	5.007	0.080	7	1950
σ_a^2		0.3	0.316	0.077	0.302	0.074	<5	20
σ_g^2		0.3	0.341	0.097	0.296	0.052	46	267

NOTE: The last two columns show Δ_{RMSE}(SG, PX − HC) and Δ_{ESS}(SG, PX − HC), respectively. For the other parameters not included both improvement measures are in the range (0, 5%).

and

$$\Delta_{ESS}(A_1, A_2) = 100 \times \left(\frac{ESS_{A_2} - ESS_{A_1}}{ESS_{A_1}} \right). \tag{4.2}$$

For the parameters that do not appear in the table, the SG sampler performs well and no improvement has been noticed. The calculations are based on 100 independent replications of the analysis under each simulation scenario. The tables show small reductions in RMSE but great improvements in ESS. For some of the parameters the effective sample size is increased more than 10-fold. The size of ESS is important when an MCMC algorithm is run until the desired Monte Carlo standard deviation is achieved for each component of the chain. In other words, our improvements in ESS show that to achieve the same degree of precision, the SG sampler must be run 10 to 20 times longer. In a genetics study in which one has to repeat the analysis for a large number of candidate SNPs the improvement makes a big practical difference.

The increase in efficiency is dramatic for the loading factors which are of direct interest in genetic studies such as the ones described in Section 1, because precise estimates of λ_j are required to detect pleiotropy. From Table 2 we find that the improvement for λ_3 and

Table 2. Simulation results under model M2: parameter estimation comparison using SG and PX^2-HC

Parameters		Value	SG		PX^2-HC			
			Est.	RMSE	Est	RMSE	Δ_{RMSE}	Δ_{ESS}
α	α_1	−1.0	−1.003	0.024	−1.002	0.024	<5	923
	α_2	1.0	1.000	0.050	1.000	0.050	<5	83
λ	λ_1	2.0	2.001	0.039	2.001	0.036	8	1124
	λ_2	3.0	3.001	0.060	3.002	0.057	5	1145
	λ_3	1.0	1.010	0.054	1.001	0.051	5	361
	λ_4	1.0	1.017	0.062	1.009	0.057	8	381
σ_a^2		1.0	1.021	0.140	1.019	0.136	<5	166
σ_g^2		1.0	1.024	0.188	1.022	0.190	<5	34

NOTE: The last two columns show Δ_{RMSE}(SG, PX^2 − HC) and Δ_{ESS}(SG, PX^2 − HC), respectively. For the other parameters not included both improvement measures are in the range (0, 5%).

λ_4, which are the factor loadings of the binary outcomes, is not as impressive as the factor loadings corresponding to the continuous outcomes. This observation is consistent with the findings of Ghosh and Dunson (2009).

The improvement in computational performance translates into more precise inference. For instance, we have investigated its impact on the coverage of the 95% highest posterior density intervals (HpdI) for the λ's in continuous and mixed models (i.e., M1 and M2) and find that the parameter expanded samples yield HpdI's with coverage rates closer to the nominal values than those constructed from SG samples. For example, under scenario M1, the empirical coverages of the 95% HpdI's for λ_1, λ_2, and λ_3 are, respectively, {88%, 89%, 86%} for SG, while for the PX-HC algorithm the coverages jump to {93%, 94%, 93%}. Illustrative plots of HpdI coverages for SG and PX algorithms under scenarios M1 and M2 are provided in the online Appendix.

Most of the SNPs considered in a genome-wide association study (GWAS) will not be associated with the LV, which means that frequency properties of Bayesian measures of significance are important in practice. We investigate the Type I error of the proposed model when using 95% highest posterior density interval (HpdI) to assess the importance of the genetic effect α_2. An effect is deemed significant if the 95% HpdI does not include zero. We have generated 100 independent replications based on scenario M2 except that α_2 is set to be zero and the value of MAF varies from 0.2 to 0.4. The empirical type I errors obtained are {0.046, 0.045, 0.051} for MAF={0.2, 0.3, 0.4}, respectively. In Figure 4, we report the 95% HpdI's from 100 replicated data under $H_0 : \alpha_2 = 0$ and the empirical coverage of $\alpha_2 = 0$ is 96%.

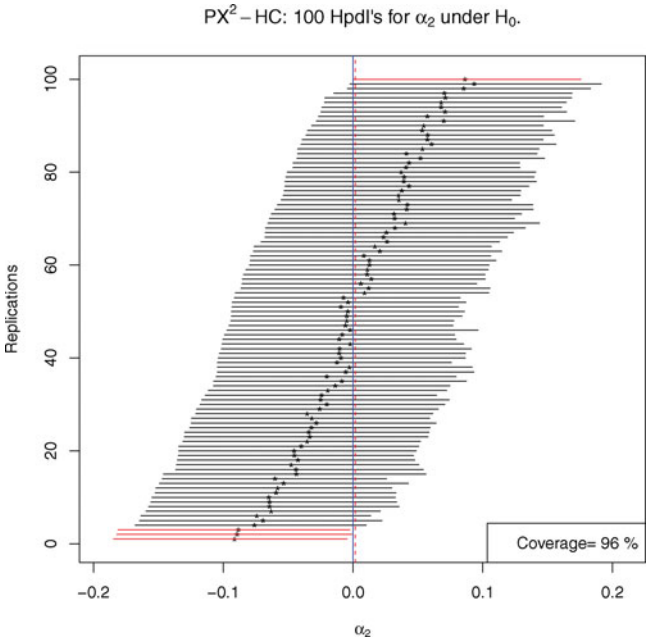

Figure 4. The 95% highest posterior density interval (HpdI) for α_2, the effect size of the genotype of a SNP with MAF = 0.3, under the null hypothesis. The HpdIs are ordered by their lower bounds. The solid vertical line is the true value, which is 0, of α_2. The dashed vertical line is the mean estimation. The empirical coverage of $\alpha = 0$ is 96%.

Table 3. Simulation results under model M2: parameter estimation comparison for $\boldsymbol{\lambda}$ and $\boldsymbol{\alpha}$ between considering and ignoring the cluster correlation existing in the data

Parameters		True	Considering cluster			Ignoring cluster		
			bias	sd	RMSE	bias	sd	RMSE
α	α_1	−1.0	0.006	0.023	0.024	0.369	0.026	0.370
	α_2	1.0	−0.004	0.046	0.046	−0.370	0.066	0.376
λ	λ_1	2.0	0.009	0.036	0.037	1.176	0.113	1.181
	λ_2	3.0	0.016	0.054	0.056	1.754	0.165	1.761
	λ_3	1.0	0.017	0.056	0.058	0.608	0.099	0.616
	λ_4	1.0	0.008	0.058	0.058	0.595	0.103	0.604

If an exceedingly large number of SNPs are investigated, one may need to use a simpler screening method to reduce the number of candidates to less than 1000. The C program used to sample from the posterior (included in the online supplemental material) requires about 2.5 min to produce 25,000 samples with 100 pedigrees included in the study. Therefore 1000 SNPs could be analyzed sequentially in two days. However, in this case a more stringent control of Type I error must be used to declare significance. Although an extended simulation study about appropriate false discovery control goes beyond the scope of this article, efficient sampling algorithms such as the one proposed here are critical to the inferential process.

4.3 EFFECT OF IGNORING FAMILY STRUCTURE

In Simulation M2, we also compare the parameter estimates obtained from the model taking into account the familial dependence with the values assuming independence. The results presented in Table 3, are in agreement with the derivations in Section 2.1 and show that naively ignoring the family structure present in the data will cause overestimation of the factor loadings and underestimation of the absolute values of the fixed effects of important covariates on the LV.

5. REAL DATA EXAMPLE: GENETIC STUDY OF TYPE 1 DIABETES (T1D) COMPLICATIONS

Here we demonstrate the practical utility of the proposed LV method by investigating the blood pressure data from a GAWS of various T1D complications (Paterson et al. 2010). The study sample consists of $n = 1302$ individuals with T1D from the Diabetes Control and Complications Trial (DCCT). Various phenotypes thought be to related to T1D complication severity, including glycosylated hemoglobin (HbA1c) and diastolic (DBP) and systolic blood pressure (SBP), were collected from each subject over the course of the DCCT. Additional covariates such as sex and body mass index (BMI) were also collected, and individuals were from two different cohorts and subjected to two treatment types (conventional vs. intensive). Over 800K SNPs were genotyped by the Illumina 1M bead chip assay for these individuals. The data can be obtained via the database of Genotypes and Phenotypes (dbGap) Authorized Access website at *https://dbgap.ncbi.nlm.nih.gov*.

Because T1D is a complex disease with various complication measures (the observed phenotypes), it is of great interest to quantify the conceptual latent complication status, as well as to understand the influencing factors (both genetic markers and clinical covariates). It is also valuable to determine if the various observed phenotypes are truly associated with the latent variable. However, previous analyses have been limited to the standard single phenotype approach in which each phenotype is analyzed separately. For example, Ye et al. (2010) performed GWAS, *separately*, for DBP and SBP, two normally distributed outcomes, and they identified rs7842868 on chromosome 8 as a SNP significantly associated with DBP. Our goal here is to formally perform a multi-phenotype analysis, jointly analyzing the measured manifest variables using the proposed Bayesian LV methodology. This approach allows us to determine if rs7842868 is associated with the latent conceptual T1D complication variable and to test if DBP and SBP are truly related to the LV. Of practical interest is whether there are other phenotypes such as hyperglycaemia (HPG) that could be included as manifest variables associated with the latent T1D severity variable. Therefore, we investigate three phenotypes, among which two are continuous (DBP and SBP) and one is binary (HPG $= 1$ for very high levels of glycaemia, that is, Hb1Ac> 8, and $= 0$ otherwise). Besides clinical considerations, this choice also allows us to evaluate the proposed method for general traits as described in Section 3.1.

All patients have consecutive quarterly visit measurements. The number of quarterly visits per patient ranges from between 2 and 10 with a median of 7 visits. Among the 1302 patients, 71 of them have less than five visits with one patient has only two visits, and the number of patients who have five to ten visits are $\{234, 337, 280, 141, 16, 223\}$, respectively. In this dataset, there is only one person in each family, therefore there is no familial correlation, but the proposed methodology can be used by assuming the cluster size is equal to 1. Since we are dealing with independent individuals, we do not include random effects a_c and g_c so the LV model becomes

$$U_i = \mathbf{X}_i\boldsymbol{\alpha} + \boldsymbol{\epsilon}_i, \ \forall 1 \leq i \leq n. \tag{5.1}$$

We first consider rs7842868, a SNP found by Ye et al. (2010) to be associated with DBP. The set of available covariates includes BMI, sex, cohort, treatment and the genotype of the SNP. We generated ten bootstrap samples using each patient as a sampling unit, and compared the DIC differences between the model that includes all the covariates in the direct effect set and the model that moves the investigated covariate to the indirect effect set. Results show that BMI has a mean increase of DIC value 302.3, while other covariates have mean decreased DIC values: $\{-0.73, -93.5, -62.3, -34.4\}$ for sex, cohort, treatment, and genotype of SNP rs7842868, respectively. Combining the suggestions from clinicians along with the DIC statistics, we assumed that the direct effect covariates W include BMI, while the indirect covariates include sex, cohort, treatment, in addition to the genotype of SNP rs7842868.

Along with the HpdI, we also calculate log Bayes factor (logBF) to test whether the factor loading λ or the indirect effect α is significant. The logBF is calculated using path sampling (PS) implemented with the parametric arithmetic mean path scheme (Lee and Song 2002). The parametric path is constructed using a scalar $s \in [0, 1]$ to link two models M_0 and M_1. For instance, to test the significance of $\lambda_{j'}$, M_0 and M_1 correspond to the models having the factor loading vectors equal to $(\lambda_1, \ldots, \lambda_{j'-1}, 0, \lambda_{j'+1}, \ldots, \lambda_J)'$ and

$(\lambda_1, \ldots \lambda_{j'-1}, \lambda_{j'}, \lambda_{j'+1} \ldots, \lambda_J)'$, respectively. The latent variable part of the model is the same as defined in Equation (5.1) for both M_0 and M_1. The two models are linked up by models M_s, $0 \leq s \leq 1$, where the factor loading vector in M_s is equal to $(\lambda_1, \ldots, \lambda_{j'-1}, s\lambda_{j'}, \lambda_{j'+1}, \ldots, \lambda_J)'$. Gelman and Meng (1998) proved that

$$\log \mathrm{BF}_{10} = \log \frac{P(Y|M_1)}{P(Y|M_0)} = \int_0^1 E_{\Omega,\Theta}[\mathbf{U}(Y, \Omega, \Theta, s)]ds, \tag{5.2}$$

where Ω is the vector of latent variables (including random effects), $E_{\Omega,\Theta}$ denotes the expectation with respect to the distribution $P(\Omega, \Theta|Y, s)$, and

$$\mathbf{U}(Y, \Omega, \Theta, s) = \frac{d}{ds} \log P(Y, \Omega|\Theta, s).$$

To evaluate the integral in Equation (5.1), we follow the procedure in Gelman and Meng (1998) and use 30 grid points ranging evenly from 0 to 1 so that $s_{(0)} = 0 < s_{(1)} < s_{(2)} < \cdots < s_{(G)} < s_{(G+1)} = 1$ and then estimate $\log \mathrm{BF}_{10}$ by

$$\widehat{\log \mathrm{BF}_{10}} = \frac{1}{2} \sum_{g=0}^{G} (s_{(g+1)} - s_{(g)}) \left(\overline{\mathbf{U}}_{(g+1)} + \overline{\mathbf{U}}_{(g)} \right), \tag{5.3}$$

where $\overline{\mathbf{U}}_{(g)}$ is the average of the values $\mathbf{U}(Y, \Omega, \Theta, s_{(g)})$ over all the MCMC samples from $p(\Omega, \Theta|Y, s_{(g)})$. To estimate $\log \mathrm{BF}_{10}$, we first run the $PX^2 - HC$ Gibbs sampling algorithm for each of the grid points and then calculate the values of the parameters under the original inference model and use them to compute $\overline{\mathbf{U}}$ (a method which is called path sampling with parameter expansion (PS-PX) by Ghosh and Dunson 2009). Folded-t priors are used for λs, induced through parameter expansion. Similar path sampling scheme is used to test the significance of the indirect effect α.

Table 4 shows that SNP rs7842868 is significantly associated with the latent variable with estimated logBF over 10 and the 95% HpdI does not cover 0. The factor loadings are equal to 7.519 for SBP and 5.709 for DBP with both estimated logBF bigger than 50, suggesting that SBP and DBP are significantly associated with the latent variable. The factor loading for HPG is only 0.019 and the estimated logBF is equal to -5.509. The evidence from both parts of the LV model show that rs7842868 has potential pleiotropic effect on the two blood pressures. Sex and cohort are also found to be significantly associated with the latent variable, but treatment is not. We then investigate rs1358030, a SNP found by Paterson et al. (2010) to be associated with HbA1c. Based on results in Table 4, there is no evidence that rs1358030 is significantly associated with the latent variable.

To further evaluate the proposed method, we simulate genotypes for two NULL SNPs that are not associated with the phenotypes of interest. One SNP has MAF equal to 0.25, the MAF of rs7842868, and the other one has MAF equal to 0.35, the MAF of rs1358030. As expected, no significant associations are detected.

We also fit the data using the Bayesian version of the model proposed by Roy and Lin (2000), extended so that it incorporates both continuous and binary outcomes. To satisfy the balanced design and equal-spaced time assumption of their model, only those outcomes from the first five consecutive quarterly visits are used, due to the high missing rate after the fifth visit. For those patients who do not have all first five visits, we assume that the

Table 4. Application results

Analysis of SNP rs7842868				
	Parameter	Estimate	95% HpdI	$\widehat{\log BF}$
rs7842868	α_1	−0.220	(−0.285, −0.154)	10.868
sex	α_2	−0.657	(−0.738, −0.572)	71.831
cohort	α_3	0.225	(0.148, 0.305)	6.840
treatment	α_4	−0.046	(−0.125, 0.037)	−3.836
Analysis of SNP rs1358030				
	Parameter	Estimate	95% HpdI	$\widehat{\log BF}$
rs1358030	α_1	0.027	(−0.047, 0.106)	−4.639
sex	α_2	−0.812	(−0.922, −0.698)	39.965
cohort	α_3	0.280	(0.168, 0.389)	8.140
treatment	α_4	−0.001	(−0.118, 0.106)	−3.269

NOTE: SNP rs7842868 was previously identified to be associated with diastolic blood pressure (DBP) and SNP rs1358030 was previously identified to be associated with HbA1c. Phenotypes of interest are DBP and systolic blood pressure (SBP), two continuous outcomes, and hyperglycemia (HPG, defined as HbA1c greater or equal to 8), a binary outcome. All phenotypes are thought to be related to Type 1 diabetes complication severity. The coefficient αs evaluate the association between the latent variable and the genetic marker and the other covariates of interest. See Section 6 and Table 1 for more details.

values are missing at random (MAR) and replace them with the means of all the other available measurements. The percentage of missing values in the first five weeks that need to be imputed is about 1.5%. The results are shown in the online Appendix. Comparing the results obtained, we notice that both models are consistent in detecting the significance of the association between the observed outcomes and the latent variables and between the covariates and the latent disease trait. However, the HpdIs of the estimated effects obtained from the proposed method are narrower than those of Roy and Lin's (2000) method. This can avoid some ambiguities in interpreting the results. For instance, in the analysis of SNP rs7842868 both models suggest that the treatment has no effect on the LV, but the HpdI intervals produced by our and Roy and Lin's (2000) methods are $(-0.125, 0.037)$ and $(-0.004, 0.263)$, respectively.

6. DISCUSSION

We considered modifications of the standard Gibbs samplers in latent variable models with mixed outcomes. The motivation is provided by genetic studies of pleiotropy, but the scope of these models is much wider. The modifications we propose are aligned with recent efforts, theoretical and computational, on improving the efficiency of Gibbs samplers through parameter expansion techniques (Hobert and Marchev 2008; Gelman et al. 2008). The modifications proposed here result in dramatic increases in effective sample sizes that can be 20 times higher than those produced by the standard algorithms for some of the models considered. Not all parameters benefit equally from the augmentation strategy proposed here, but for the genetic pleiotropy analyses that motivated this study the improvements brought by the new algorithm in estimating the regression coefficients of genotype (α) and

the factor loadings (λ) are of great importance. While we have not registered losses in Monte Carlo efficiency or effective sample size, we would like to continue searching for other auxiliary variable constructs that will impact all the parameters in the model. These efforts will be reported in future communications.

We have not expanded on issues related to model selection and variable selection. In Section 5 we have relied on Bayes factors to base our inclusion/exclusion of a variable in/from the model. Alternative approaches, explored by Xu (2012), include spike-and-slab priors for the parameters in focus (e.g., λ's and α's). The parameter expansion approach can be directly implemented in such setups with the folded-t prior distributions being replaced by mixtures of folded-t and point mass distributions.

The computational load of the proposed algorithms is still too high to perform a genome-wide search for pleiotropic genetic variants. The recent advances in parallel computing can partially alleviate this constraint. Alternatively, a two-stage approach can be used in which a simple and less stringent selection procedure is first used to select a moderate number of candidate variants for further investigation using the proposed method. The uncertainty inherited from the first-stage selection, however, must be accounted for in the models used in the second stage.

SUPPLEMENTARY MATERIALS

Supplementary materials for this article are available on the publisher's website.

ACKNOWLEDGMENTS

The authors thank the Editor, Thomas Lee, the Associate Editor, and two anonymous referees for their suggestions that have greatly improved the article. This work was supported by the Natural Sciences and Engineering Research Council (NSERC) of Canada and the Canadian Institutes of Health Research (CIHR; MOP 84287) to RVC and LS, the Ontario Graduate Scholarship (OGS) to LX. ADP holds a Canada Research Chair in the Genetics of Complex Diseases and received funding from Genome Canada through the Ontario Genomics Institute.

The DCCT/EDIC Research Group is sponsored through research contracts from the National Institute of Diabetes, Endocrinology, and Metabolic Diseases of the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) and the National Institutes of Health (N01-DK-6-2204, R01-DK-077510). The Diabetes Control and Complications Trial (DCCT) and its follow-up the Epidemiology of Diabetes Interventions and Complications (EDIC) study were conducted by the DCCT/EDIC Research Group and supported by National Institute of Health grants and contracts and by the General Clinical Research Center Program, NCRR. The data and samples from the DCCT/EDIC study were supplied by the NIDDK Central Repositories. This manuscript was not prepared under the auspices of and does not represent analyses or conclusions of the NIDDK Central Repositories, or the NIH.

[Received July 2013. Revised November 2014.]

REFERENCES

Albert, J., and Chib, S. (1993), "Bayesian Analysis of Binary and Polychotomous Response Data," *Journal of the American Statistical Association*, 88, 669–679. [412]

Bartholomew, D., Knott, M., and Moustaki, I. (2011), *Latent Variable Models and Factor Analysis: A Unified Approach*, Wiley Series in Probability and Statistics, New York: Wiley. [405]

Burton, P., Scurrah, K., Tobin, M. D., and Palmer, L. (2005), "Covariance Components Models for Longitudinal Family Data," *International Journal of Epidemiology*, 34, 1063–1067. [406]

Diggle, P. J., Liang, K.-Y., and Zeger, S. L. (1994), *Analysis of Longitudinal Data* (1st ed.), London: Oxford Science. [408]

Engle, R. W., Tuholski, S. W., Laughlin, J. E., and Conway, A. R. A. (1999), "Working Memory, Short-Term Memory, and General Fluid Intelligence: A Latent-Variable Approach," *Journal of Experimental Psychology*, 128, 309–331. [406]

Flegal, J., Haran, M., and Jones, G. (2008), "Markov Chain Monte Marlo: Can We Trust the Third Significant Figure?," *Statistical Science*, 23, 250–260. [415]

Gelfand, A. E. (1995), "Efficient Parametrisations for Normal Linear Mixed Models," *Biometrika*, 82, 479–488. [406,410]

——— (2006), "Prior Distributions for Variance Parameters in Hierarchical Models," *Bayesian Analysis*, 1, 515–533. [411]

Gelman, A., and Meng, X.-L. (1998), "Simulating Normalizing Constants: From Importance Sampling to Bridge Sampling to Path Sampling," *Statistical Science*, 13, 163–185. [421]

Gelman, A., and Rubin, D. B. (1992), "Inference from Iterative Simulation Using Multiple Sequences" (with discussion), *Statistical Science*, 457–511. [416]

Gelman, A., van Dyk, D., Huang, Z., and Boscardin, J. (2008), "Using Redundant Parametrizations to Fit Hierarchical Models," *Journal of Computational and Graphical Statistics*, 17, 95–122. [406,411,422]

Geyer, C. J. (1992), "Practical Markov Chain Monte Carlo" (with discussion), *Statistical Science*, 7, 473–483. [415]

Ghosh, J., and Dunson, D. B. (2009), "Default Prior Distributions and Efficient Posterior Computation in Bayesian Factor Analysis," *Journal of Computational and Graphical Statistics*, 18, 306–320. [411,418,421]

Hobert, J. P., and Marchev, D. (2008), "A Theoretical Comparison of the Data Augmentation, Marginal Augmentation and PX-DA Algorithms," *Annals of Statistics*, 36, 532–554. [406,422]

Jamsen, K. M., Zaloumis, S. G., Scurrah, K. J., and Gurrin, L. C. (2010), "Specification of Generalized Linear Mixed Models for Family Data using Markov Chain Monte Carlo Methods," *Journal of Biometrics and Biostatistics*, S1–003. [407]

Jiang, D., and McPeek, M. S. (2014), "Robust Rare Variant Association Testing for Quantitative Traits in Samples With Related Individuals," *Genetic Epidemiology*, 38, 10–20. [413]

Khatab, K., and Fahrmeir, L. (2009), "Analysis of Childhood Morbidity With Geoadditive Probit and Latent Variable Model: A Case Study for Egypt," *The American Journal of Tropical Medicine and Hygiene*, 81, 116–128. [408]

Kuttner, K. N. (1994), "Estimating Potential Output as a Latent Variable," *Journal of Business and Economic Studies*, 12, 361–368. [406]

Lee, S. Y., and Song, X. Y. (2002), "Bayesian Selection on the Number of Factors in a Factor Analysis Model," *Behaviormetrika*, 29, 23–40. [420]

Liu, J. S., and Wu, Y. N. (1999), "Parameter Expansion for Data Augmentation," *Journal of the American Statistical Association*, 94, 1264–1274. [406,410]

Meng, X.-L., and van Dyk, D. (1999), "Seeking Efficient Data Augmentation Schemes via Conditional and Marginal Augmentation," *Biometrika*, 86, 301–320. [406,410]

Paterson, A. D., Waggott, D., Boright, A. P, Hosseini, S. M., Shen, E., Sylvestre, M.-P., Wong, I., Bharaj, B., Cleary, P. A., Lachin, J. M., MAGIC (Meta-Analyses of Glucose and Insulin-related traits Consortium), Below, J. E., Nicolae, D., Cox, N. J., Canty, A. J., Sun, L., Bull, S. B. and the Diabetes Control and Complications Trial/Epidemiology of Diabetes Interventions and Complications Research Group (2010), "A Genome-Wide Association Study Identifies a Novel Major Locus for Glycemic Control in Type 1 Diabetes, as Measured by Both A1C and Glucose," *Diabetes*, 59, 539–549. [419,421]

Roy, J., and Lin, X. (2000), "Latent Variable Models for Longitudinal Data With Multiple Continuous Outcomes," *Biometrics*, 56, 1047–1054. [405,406,409,421]

Sammel, M. D., and Ryan, L. M. (1996), "Latent Variable Models with Fixed Effects," *Biometrics*, 52, 650–663. [405,408]

——— (1997), "Latent Variable Models for Mixed Discrete and Continuous Outcomes," *Journal of the Royal Statistical Society,* Series B, 59, 667–678. [405]

Sanchez, B., Budtz-Jorgensen, E., Ryan, L., and Hu, H. (2005), "Structural Equation Modeling: A Review with Applications to Environmental Epidemiology," *Journal of the American Statistical Association*, 100, 1443–1455. [406]

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. [409]

Tanner, M., and Wong, W. (1987), "The Calculation of Posterior Distributions by Data Augmentation," *Journal of the American Statistical Association*, 82, 528–540. [410]

Thornton, T., and McPeek, M. S. (2010), "ROADTRIPS: Case-Control Association Testing With Partially or Completely Unknown Population and Pedigree Structure," *The American Journal of Human Genetics*, 86, 172–184. [409]

van Dyk, D. A., and Meng, X.-L. (2001), "The Art of Data Augmentation," *Journal of Computational and Graphical Statistics*, 10, 1–50. [413]

Xu, L. (2012), "Bayesian Methods for Genetic Association Studies," unpublished PhD thesis, University of Toronto. [409,423]

Ye, C., Canty, A. J., Waggott, D., Sylvestre, M.-P., Shen, E., Hosseini, M., et al. (2010), "A Repeated Measures Genome wide Association Study of Blood Pressure in Type 1 Diabetes," Abstract # 203 Presented at the Nineteenth Annual Meeting of the International Genetic Epidemiology Society, *Genetic Epidemiology*, 34, 973. [420]