Latent variable modeling paradigms for genotype-trait association studies

Yan Liu and **Andrea S. Foulkes***

Division of Biostatistics, 404 Arnold House, 715 North Pleasant Street, Amherst, MA 01003, USA

Received 15 October 2010, revised 14 June 2011, accepted 27 June 2011

Characterizing associations among multiple single-nucleotide polymorphisms (SNPs) within and across genes, and measures of disease progression or disease status will potentially offer new insight into disease etiology and disease progression. However, this presents a significant analytic challenge due to the existence of multiple potentially informative genetic loci, as well as environmental and demographic factors, and the generally uncharacterized and complex relationships among them. Latent variable modeling approaches offer a natural framework for analysis of data arising from these population-based genetic association investigations of complex diseases as they are well-suited to uncover simultaneous effects of multiple markers. In this manuscript we describe application and performance of two such latent variable methods, namely structural equation models (SEMs) and mixed effects models (MEMs), and highlight their theoretical overlap. The relative advantages of each paradigm are investigated through simulation studies and, finally, an application to data arising from a study of antiretroviral-associated dyslipidemia in HIV-infected individuals is provided for illustration.

Keywords: Mixed effects models; Single-nucleotide polymorphisms (SNPs); Structural equation model (SEM).

Supporting Information for this article is available from the author or on the WWW under http://dx.doi.org/10.1002/bimj.201000218.

1 Introduction

The increased availability of data on single-nucleotide polymorphisms (SNPs) has led to heighten interest in understanding how this genetic information correlates with measures of disease progression. One analytic challenge plaguing these genotype-trait association studies is the potential for multiple SNPs to be implicated in complex diseases. In this manuscript, we describe applications and performance of two latent variable paradigms, namely structural equation models (SEMs) and mixed effects models (MEMs), for addressing this challenge.

SEMs constitute a broad range of multivariate regression models that allow complex dependencies among multiple predictors and outcome variables and are widely used in economics, sociology and psychology (Pugesek et al., 2003; Rabe-Hesketh et al., 2004; Skrondal and Rabe-Hesketh, 2004). Several recent manuscripts extend the conventional measurement component of an SEM, conditional on latent variables, to the generalized linear model setting, rendering these models naturally conducive to continuous as well as categorical outcomes (Muthén, 1984; Muthén and Muthén, 2007; Skrondal and Rabe-Hesketh, 2005, 2004; Lee and Shi, 2001; Reboussin and Liang, 1998). Recent applications of SEMs to genetic data include those that aim to reconstruct the linkage disequilibrium structure among genes (Lee et al., 2007) as well as one study to characterize

*Corresponding author: e-mail: foulkes@schoolph.umass.edu, Phone: +1-413-545-1881, Fax: +1-413-545-1645

associations between multiple SNPs, smoking, gender and rheumatoid arthritis (Nock et al., 2007). MEMs, widely used to address correlations in repeated-measures and multi-level data (Laird and Ware, 1982), are an alternative latent variable modeling strategy that has been described for characterizing association between multiple SNPs, within and across genes, and a measured trait (Foulkes et al., 2005; Goeman et al., 2004; Foulkes and De Gruttola, 2002).

A growing body of literature exists on the methods for analyzing data arising from candidate gene association studies, including approaches targeted specifically at characterizing combinations of SNPs and their association with a measure of disease status or disease progression. Among these are most notably machine learning applications, including classification and regression trees (CART) (Zhang and Singer, 1999; Breiman et al., 1993), random forests (Bureau et al., 2005; Segal et al., 2004; Breiman, 2001), logic regression (Schwender and Ickstadt, 2008; Kooperberg and Ruczinski, 2005; Ruczinski et al., 2004, 2003; Kooperberg et al., 2001), lasso (Kooperberg et al., 2010; Wu et al., 2009; Tibshirani, 1996), elastic net (Kooperberg et al., 2010; Zou and Hastie, 2005) and Bayesian network (BN) analysis (Rodin and Boerwinkle, 2005; Pearl, 1988). The gains attributable to first-stage creation of meta-variables within these frameworks are also described, for example in Foulkes et al. (2004); Bastone et al. (2004) and Malovini et al. (2009). The former involves a first-stage, unsupervised clustering of individuals based solely on genotype data, followed by application of CART to characterize association, while the later involves a first-stage application of CART to identify clusters, followed by application of BN analysis to characterize association. The latent class approaches described herein similarly involve defining group indicators based on a collection of SNPs and, in turn, relating these to a measured trait for characterizing association; however, both the SEM and MEM approaches detailed below are distinct in that they involve fully parametric modeling of association and corresponding parameter estimation and testing. The present manuscript focuses on the overlap of these two specific latent class paradigms while additional details on several of the alternative approaches listed above, including discussion of their relative merits, can be found in Hastie et al. (2001); Gentleman et al. (2005); Schwender et al. (2008) and Foulkes (2009).

We begin by formalizing the SEM approach for genetic association studies and extend the research of Nock et al. (2007), to characterize broadly the performance under a range of underlying models of association (Section 2.1). Second, we present an extension of the MEM approach of Foulkes et al. (2005), for this setting that offers additional flexibility in defining the model of association through inclusion of cross-classified clusters (Section 2.2). We then highlight the theoretical overlap between SEMs and MEMs (Section 2.3) and explore, through simulation studies, the relative advantages of each approach (Section 3.1). Specifically, we focus on the flexibility and performance under model misspecification. Finally, we apply both approaches, as we as an alternative two-stage BN analysis, to data arising from a study of anti-retroviral therapy (ART)-associated dyslipidemia in HIV (Section 3.2).

2 Methods

2.1 Structural equation model for genetic association studies

We begin by describing how the SEM framework can be applied for analysis of data derived from genetic association studies, where the goal is to characterize associations between genotypes, within and across multiple genetic loci, and a single measure of disease progression or disease status. An extensive literature exists on SEMs, and correspondingly a variety of approaches to specifying the model have been described (Jöreskog, 1975; Bentler and Weeks, 1980; Muthén, 1984, 2002; Sánchez et al., 2005; Skrondal and Rabe-Hesketh, 2005; Muthén and Muthén, 2007). Here, we use the formulation given by Sánchez et al. (2005) and apply the measurement model described by Muthén (1984); Skrondal and Rabe-Hesketh (2005); and Muthén and Muthén (2007).

Let y_i denote the trait under study, where $i = 1, \ldots, N$ represents individuals. Further suppose $\mathbf{X}_i = (x_{i1}, \ldots, x_{iS})^{\mathrm{T}}$ represents genotypes across S bi-allelic SNPs for individual i. Since each SNP is bi-allelic, we have $x_{is} \in \{AA, Aa, aa\}$, where A and a represent the two possible nucleotides at site s and $s = 1, \ldots, S$. For simplicity of presentation, we drop the notational dependency of A and a on s. Finally, let $\mathbf{Z}_i = (z_{i1}, \ldots, z_{iP})^{\mathrm{T}}$ represent P measured covariates for individual i.

Similar to the approach described by Nock et al. (2007), we assume that each candidate gene has a corresponding latent variable, representing an unobservable effect of the corresponding gene on the trait. These latent variables are given for individual i by the vector $\mathbf{U}_i = (u_{i1}, \ldots, u_{iK})^{\mathrm{T}}$ where u_{ik} corresponds to gene k, $k = 1, \ldots, K$. The measurement component of an SEM relates the observed data components, $\mathbf{X}_i$ and y_i, to the latent variables, $\mathbf{U}_i$, while the structural component defines the relationship among the latent variables. These are formulated as follows:

$$\text{(Measurement Model)}: \quad g(\mathrm{E}(\mathbf{X}_i|\mathbf{U}_i)) = \nu_x + \Lambda_x \mathbf{U}_i + \Gamma_x \mathbf{Z}_i, \tag{1}$$

$$y_i = \nu_y + \Lambda_y \mathbf{U}_i + \Gamma_y \mathbf{Z}_i + \varepsilon_i, \tag{2}$$

where $\nu_x, \nu_y, \Lambda_x, \Lambda_y, \Gamma_x$ and Γ_y are unknown parameters, and $\varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma_e^2)$. Here $g(\cdot)$ is used to represent an appropriately defined link function, such as the probit or logit link for categorical and binary outcomes, respectively

$$\text{(Structural Model)}: \quad \mathbf{U}_i = \alpha + \mathbf{B}\mathbf{U}_i + \zeta_i, \tag{3}$$

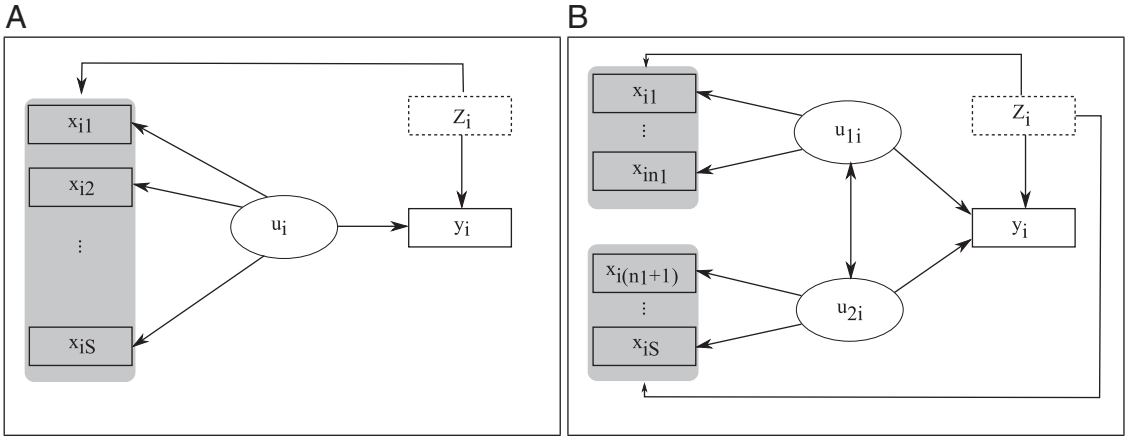
where α, $\mathbf{B}$ are unknown parameters, the diagonal elements of $\mathbf{B}$ are identically equal to 0 and $(I - \mathbf{B})$ is invertible (Sánchez et al., 2005). Here we assume $\zeta_i \sim \mathrm{MVN}(\mathbf{0}, \Psi)$ and ζ_i is independent of ε_i. In the genetic association study setting, interest is in characterizing the association between the latent variables, $\mathbf{U}_i$, and the measured trait, given by y_i. Formally, a test of association is given by a Wald test of the null hypothesis, $H_0 : \Lambda_y = 0$.

Notably, in the genetic association setting, where $\mathbf{X}_i$ represents SNPs as described above, many of the covariates represented in $\mathbf{Z}_i$ will influence the trait y_i but will not be directly predictive of $\mathbf{X}_i$ as described in Eq. (1). Covariates that are potentially relevant in this component of the measure model include surrogates for population substructure, such as race and ethnicity, as well as measures of exposure to environmental toxins, such as smoking status, that may result in oncogenic mutations within specific organ tissue. In this sense, $\mathbf{Z}_i$ can be thought of as a partitioned matrix, given by $\mathbf{Z}_i^{\mathrm{T}} = [\mathbf{Z}_{1i}^{\mathrm{T}} \quad \mathbf{Z}_{2i}^{\mathrm{T}}]^{\mathrm{T}}$, where only the covariates represented in $\mathbf{Z}_{1i}$ are potentially predictive of $\mathbf{X}_i$ while the covariates given in both $\mathbf{Z}_{1i}$ and $\mathbf{Z}_{2i}$ are potentially predictive of y_i. In turn, the element of Γ_x corresponding to $\mathbf{Z}_{2i}$ is identically equal to 0.

Visual path diagram representations of this model with one and two latent variables are given in Fig. 1A and B, respectively. Here, observed variables are represented by rectangles while latent variables are given by ovals. Dashed lines represent fixed, independent variables, whereas solid lines indicate dependent variables with corresponding distributional assumptions. Single-direction arrows represent causal relationships among variables while double-headed arrows represent non-zero correlations.

2.2 Mixed effect model for genetic association studies

Distinct from the SEM setting, application of an MEM to SNP data is a staged approach that involves first assigning individuals to groups based on observed genotypes across multiple SNPs. These genotype group assignments are then treated as cluster indicators in the usual mixed modeling framework. While several approaches to the first-stage dimension reduction are tenable, such as hierarchical or K-means clustering (Hartigan, 1975), here we apply the simple deterministic approach of assigning individuals with identical multi-locus genotypes to the same genotype group, as described by Foulkes et al. (2004).



Single latent variable SEM. SEM with two latent variables.

Figure 1 Sample SEM path diagrams for genetic association studies. (A) Single latent variable SEM and (B) SEM with two latent variables.

Again, we begin by letting $\mathbf{X}_i = (x_{i1}, \ldots, x_{iS})^{\mathrm{T}}$ represent the multilocus genotype for individual i across S bi-allelic SNPs. Now suppose $\mathbf{g} = \{g_1, \ldots, g_J\}$ represents the J groups resulting from assigning individuals with identical genotypes to the same group, that is $i, i' \in g_j$ implies $\mathbf{X}_i = \mathbf{X}_{i'}$. The MEM as described previously for this setting (Foulkes et al., 2005) can be formulated as follows:

$$y_i = \nu + \mathbf{C}_i^{\mathrm{T}}\mathbf{b} + \Gamma\mathbf{Z}_i + \varepsilon_i, \tag{4}$$

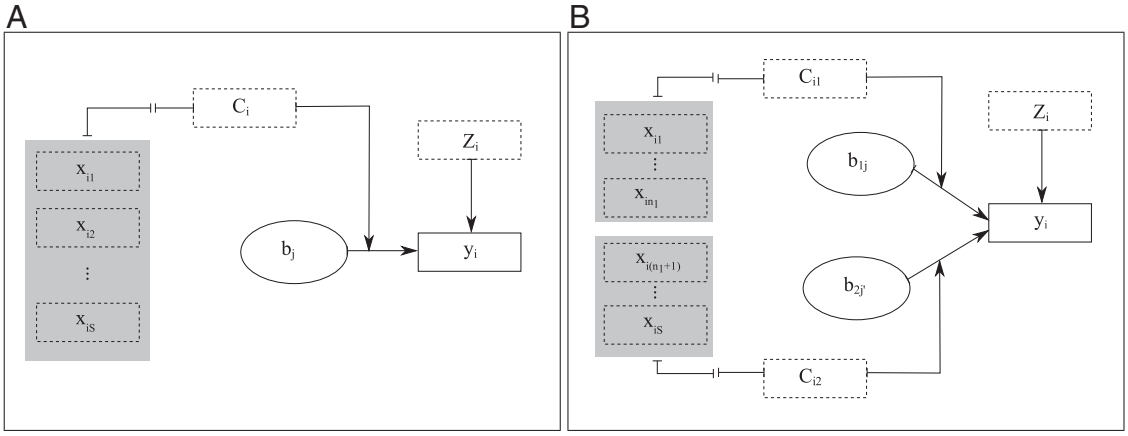
where ν and Γ are again unknown parameters, $\mathbf{C}_i = (c_{i1}, \ldots, c_{iJ})^{\mathrm{T}}$, $c_{ij} = I_{i \in g_j}$ is an indicator for individual i belonging to genotype group g_j, $\mathbf{b} = (b_1, \ldots, b_J)^{\mathrm{T}}$ is a vector of corresponding random effects of genotype groups on the trait under study, $b_j \overset{iid}{\sim} N(0, \sigma_b^2)$ for $j = 1, \ldots, J$, $\varepsilon_i \overset{iid}{\sim} N(0, \sigma_e^2)$ and b_j and ε_i are independent. A likelihood ratio test of the null hypothesis H_0: $\sigma_b = 0$ is applied to assess the presence of a genotype-trait association.

More generally, a grouping variable can be defined for each of multiple genes. To see this, suppose now that $\mathbf{X}_i$ represents a vector of S SNPs across K genes. We assume n_k SNPs are measured within gene k, such that $\sum_k n_k = S$. Now $\mathbf{g}_k = \{g_1, \ldots, g_{J_k}\}_k$ represents the groups corresponding to gene k where J_k is the number of such groups. Notably, in the setting of three-level SNPs, we have $J_k \leq 3^{n_k}$, while for binary SNPs, $J_k \leq 2^{n_k}$. The MEM for such cross-classified data is then given by

$$y_i = \nu + \sum_{k=1}^{K} \mathbf{C}_{ik}^{\mathrm{T}}\, \mathbf{b}_k + \Gamma\mathbf{Z}_i + \varepsilon_i, \tag{5}$$

where $\mathbf{C}_{ik} = (c_{i1}, , c_{iJ_k})_k^{\mathrm{T}}$ is a vector of group membership indicators, $\mathbf{b}_k = (b_{k1}, \ldots, b_{kJ_k})^{\mathrm{T}}$ is defined as the genotype group random effects on y_i for gene k, $b_{kj} \sim N(0, \sigma_{bk}^2)$ and $\varepsilon_i \overset{iid}{\sim} N(0, \sigma_e^2)$. In this setting, a likelihood ratio test can again be applied to test the null hypothesis $H_0 : \sigma_{bk} = 0$ for each gene k.

Visual representations of the MEMs for single- and multi-level clustering are given in Fig. 2A and B, using the same notation as described above for Fig. 1. Here, the broken lines indicate a deterministic relationship between SNPs – represented by $x_{i1}, \ldots, x_{iS}$ – and cluster assignments – represented by $\mathbf{C}_i$. A few notable distinctions can be discerned by comparing Figs. 1 and 2. Most notably, in the SEM framework, we see that the SNP variables are treated as random, and modeled as a function of the latent variables, u_i. In the MEM setting, on the other hand, these are treated as fixed and inform cluster assignments deterministically. Additionally, in the SEM setting, the latent



Single level clustering MEM. MEM with multi-level clustering.

Figure 2 Sample MEM path diagrams for genetic association studies. (A) Single level clustering MEM and (B) MEM with multi-level clustering.

variables – given by u_{i1} and u_{i2} – are person-specific and potentially correlated, while in the MEM framework the latent effects – given by b_{1j} and $b_{2j'}$ – correspond to genotype groups and are independent. Further discussion of theoretical overlap between the two modeling approaches is given in Section 2.3.

2.3 A comparison of the SEM and MEM approaches

Both the SEM and MEM approaches, as formulated in Sections 2.1 and 2.2, involve modeling underlying latent variables that represent unobservable effects of genes on the trait under study. Indeed, it is straightforward to show that the SEM can be reduced to an MEM for this setting. To begin, consider the simple case of a single gene, and thus a single latent variable. We first omit the regression of $\mathbf{X}_i$ on the latent variable – Eq. (1) of the SEM – as the MEM treats the $\mathbf{X}_i$ as fixed. For the single gene setting, the models of Eqs. (2) and (3) reduce to

$$y_i = \nu_y + \lambda_y u_i + \Gamma_y \mathbf{Z}_i + \varepsilon_i, \tag{6}$$

$$u_i = \alpha + \zeta_i, \tag{7}$$

where $\Lambda_y = \lambda_y$ is now a scalar and, because there is only one assumed latent variable, $\mathbf{B}$ is identically equal to 0 and $\zeta_i \overset{iid}{\sim} N(0, \psi)$. In order for this SEM to reduce to the MEM, we let λ_y equal the vector $\mathbf{C}_i^{\mathrm{T}}$ and replace the individual level latent variable u_i with the vector of random cluster effects $\mathbf{b} = (b_1, \ldots, b_J)^{\mathrm{T}}$. Importantly, this is equivalent to making the assumption that the latent effects on the trait are the same for those individuals with the same observed genetic profile. Finally, we set $\alpha = 0$ and, together, these restrictions yield Eq. (4).

In the case of $K = 2$ genes, we note that Eqs. (2) and (3) can be written as:

$$y_i = \nu_y + (\lambda_{y1} \quad \lambda_{y2}) \begin{pmatrix} u_{1i} \\ u_{2i} \end{pmatrix} + \Gamma_y \mathbf{Z}_i + \varepsilon_i, \tag{8}$$

$$\begin{pmatrix} u_{1i} \\ u_{2i} \end{pmatrix} = \alpha + \begin{bmatrix} 0 & B_{12} \\ B_{21} & 0 \end{bmatrix} \begin{pmatrix} u_{1i} \\ u_{2i} \end{pmatrix} + \zeta_i. \tag{9}$$

Now we replace λ_{y1} and λ_{y2} with the vectors $\mathbf{C}_{i1}$ and $\mathbf{C}_{i2}$ and replace u_{1i} and u_{2i} with $\mathbf{b}_1^{\mathrm{T}}$ and $\mathbf{b}_2^{\mathrm{T}}$, respectively. Notably, as the lengths of $\mathbf{C}_{i1}$ and $\mathbf{C}_{i2}$ (as well as $\mathbf{b}_1$ and $\mathbf{b}_2$), given by J_1 and J_2, are not

necessarily equal, these vectors need to be concatenated with vectors of 0 of appropriate length. That is, Eq. (8) is replaced with:

$$y_i = \nu_y + \left[\begin{pmatrix} \mathbf{C}_{i1} \\ \underset{\sim}{\mathbf{0}} \end{pmatrix} \quad \mathbf{C}_{i2} \right] \begin{bmatrix} (\mathbf{b}_1^{\mathrm{T}} & \underset{\sim}{\mathbf{0}}) \\ \mathbf{b}_2^{\mathrm{T}} & \end{bmatrix} + \Gamma_y \mathbf{Z}_i + \varepsilon_i$$

where, without loss of generality, we assume $J_1 < J_2$ and $\underset{\sim}{\mathbf{0}}$ is defined as a vector of length $(J_2 - J_1)$ with all 0 elements. In order for this SEM to simplify to the MEM with two levels of clustering, we additionally need to assume $\alpha = 0$ and $B_{12} = B_{21} = 0$. In other words, we must assume that the latent variables are uncorrelated and centered at 0.

In summary, and more generally for $K > 2$ genes, we make the following three assumptions for the SEM to reduce to the MEM: (i) SNPs, represented by $\mathbf{X}_i$, are fixed, so that the model of Eq. (1) is omitted; (ii) $\Lambda_y \mathbf{U}_i$ is given by $\mathbf{C}_i \mathbf{b}$, that is individual level latent variables are the same for individuals within the same defined genotype group and (iii) $\alpha = \mathbf{B} = 0$, that is latent variables across genes are mutually independent and centered at 0. In the MEM setting, the cluster random effects are assumed independent, although a correlation structure between $\mathbf{b}_j$ and $\mathbf{b}_{j'}$ could be imposed.

3 Applications

In the following sections we report the results of a simulation study and an application to a study of anti-retroviral-associated dyslipidemia in HIV. Restricted maximum likelihood is used to derive point estimates of parameters in the MEMs. A likelihood ratio test of $H_0 : \sigma_b = 0$, comparing the full (mixed effects) and reduced (fixed effects only) model, is used to investigate the association between genotype groups and a measured trait. As this involves testing a parameter at a boundary, a mixture of a χ^2 with 1 and 0 degrees of freedom is assumed for the resultant test statistic. All MEMs are fitted with the `lme()` function within the `nlme` package in `R`, Version 2.9.1. In the context of fitting SEMs, weighted least squares is applied to derive parameter estimates and a Wald test of the null hypothesis $H_0 : \lambda_y = 0$, is reported. SEMs are fitted using `Mplus` Version 5.21.

3.1 Simulation studies

In this section we explore, through simulation studies, the relative practical performance of SEMs and MEMs under a range of underlying models of association. In each simulation, we generate 500 sets of data, each of size $n = 1000$, for each combination of true parameter values. SEM data are simulated using the `MONTECARLO` Command in `Mplus` Version 5.21 (Muthén, 1984; Muthén and Muthén, 2007).

We begin by simulating data under an SEM with a single latent variable, according to Eqs. (1) and (6)–(7), where for simplicity of presentation, we let $\Gamma_x = 0$. In scalar notation, this model can be rewritten as:

$$\begin{aligned} g(\mathrm{E}(x_{is}|u_i)) &= \nu_{xs} + \lambda_{xs} u_i, \\ y_i &= \nu_y + \lambda_y u_i + \gamma_y z_i + \varepsilon_i, \\ u_i &= \alpha + \zeta_i \end{aligned}$$

where we assume $s = 1, \ldots, 4$, x_{is} is a binary SNP indicator, $z_i \sim N(0,1)$, $\varepsilon_i \sim N(0, \sigma_e^2)$, $\zeta_i \sim N(0, \psi)$ and $\varepsilon_i \perp \zeta_i$. For identifiability, α is set to 0 and λ_{x1} is restricted to 1. Furthermore, we define a threshold model in `Mplus` such that $\mathrm{P}(x_{is} = 1) = 0.50$. It is straightforward to show that the covariance between any two SNPs is then given by ψ. Finally, we set $\nu_{xs} = 0$, $\lambda_{x2} = \lambda_{x3} = \lambda_{x4} = 1$ and $\nu_y = 0$, while the values of the remaining parameters, λ_y, γ_y, σ_e^2, and ψ, vary as described in Table 1.

Table 1 Simulation results under SEM with one latent variable.

Data	True value	SEMs				MEMs			
		Bias	CR	CI	Power	Bias	CR	CI	Power
1	$\sigma_e^2 = 1$	0.00	0.95	0.23	–	0.13	0.24	0.20	–
	$\psi = 0.2$	0.01	0.96	0.21	–	–	–	–	–
	$\gamma_y = 1$	0.00	0.95	0.14	–	0.00	0.96	0.13	–
	$\lambda_y = 1$	−0.02	0.94	0.73	1.00	–	–	–	1.00
2	$\sigma_e^2 = 1$	0.00	0.95	0.23	–	0.18	0.06	0.21	–
	$\psi = 0.4$	0.01	0.96	0.20	–	–	–	–	–
	$\gamma_y = 1$	0.00	0.95	0.15	–	0.00	0.96	0.14	–
	$\lambda_y = 1$	−0.01	0.95	0.37	1.00	–	–	–	1.00
3	$\sigma_e^2 = 1$	0.00	0.94	0.22	–	0.19	0.03	0.21	–
	$\psi = 0.6$	0.01	0.96	0.17	–	–	–	–	–
	$\gamma_y = 1$	0.00	0.95	0.16	–	0.00	0.95	0.14	–
	$\lambda_y = 1$	0.00	0.97	0.25	1.00	–	–	–	1.00
4	$\sigma_e^2 = 1$	0.00	0.94	0.19	–	0.04	0.85	0.18	–
	$\psi = 0.4$	0.01	0.95	0.22	–	–	–	–	–
	$\gamma_y = 1$	0.00	0.96	0.13	–	0.00	0.96	0.13	–
	$\lambda_y = 0.5$	−0.01	0.96	0.30	1.00	–	–	–	1.00
5	$\sigma_e^2 = 1$	0.01	0.95	0.39	–	0.69	0.00	0.30	–
	$\psi = 0.4$	0.00	0.96	0.18	–	–	–	–	–
	$\gamma_y = 1$	0.00	0.95	0.20	–	0.00	0.95	0.16	–
	$\lambda_y = 2$	−0.02	0.95	0.56	1.00	–	–	–	1.00

(a) Median estimates are reported from 500 sets of data. (b) Absolute difference between Est. and true value. (c) Coverage rate, the percentage of confidence intervals (CIs) that cover true value among 500 CIs. For the CR of variance, the CIs that contain negative values are excluded for consideration. (d) Median length among the 500 length of CIs. (e) Wald test statistics of $H_0 : \lambda_y = 0$ is used to test the association between latent variable and measured trait. (f) Likelihood ratio test (LRT) of $H_0 : \sigma_b = 0$ is applied to investigate the association between genotype groups and measured trait. Here the approximate distribution of LRT under null hypothesis is 50:50 mixture of χ_0^2 and χ_1^2.

The results of fitting both SEMs and MEMs to these data are provided in Table 1. In general, both approaches have high power for detecting association, but the SEM approach performs better in terms of bias and coverage for the variance component, σ_e^2. As ψ and λ_y increase while σ_e^2 remains fixed, the absolute bias associated with σ_e^2 for the MEM increases and the corresponding coverage rate (CR) is low. Notably, the type-1 error rate (under the model in which $\lambda_y = 0$) is 0.03 for both the SEM and MEM approaches.

Second, we simulate data according to an SEM with two uncorrelated latent variables. That is, we let the data arise from Eqs. (1) and (8)–(9) with $B = 0$. In this case, our model can be written as:

$$
\begin{aligned}
g(\mathrm{E}(x_{is}|u_{1i})) &= \nu_{xs} + \lambda_{xs} u_{1i}, \\
g(\mathrm{E}(x_{it}|u_{2i})) &= \nu_{xt} + \lambda_{xt} u_{2i}, \\
y_i &= \nu_y + \lambda_{y1} u_{1i} + \lambda_{y2} u_{2i} + \gamma_y z_i + \varepsilon_i, \\
u_{1i} &= \alpha_1 + \zeta_{1i}, \\
u_{2i} &= \alpha_2 + \zeta_{2i},
\end{aligned}
$$

where $s = 1, \ldots, 4$ and $t = 5, \ldots, 8$, that is we have 8 SNPs with 4 in each of two genes. We further assume $\varepsilon_i \sim N(0, \sigma_e^2)$, $\zeta_{1i} \sim N(0, \psi_1)$, $\zeta_{2i} \sim N(0, \psi_2)$, all mutually independent and $z_i \sim N(0, 1)$. For identification, α_1 and α_2 are set to 0 and we restrict $\lambda_{x1} = \lambda_{x5} = 1$. For the simulation, we set

Table 2 Simulation results under SEM with two uncorrelated latent variables.

Data	True value	SEMs				MEMs			
		Bias	CR	CI	Power	Bias	CR	CI	Power
1	$\sigma_e^2 = 1$	−0.01	0.94	0.30	–	0.25	0.00	0.22	–
	$\psi_1 = 0.2$	0.00	0.95	0.21	–	–	–	–	–
	$\psi_2 = 0.2$	0.00	0.94	0.21	–	–	–	–	–
	$\gamma_y = 1$	0.00	0.94	0.15	–	0.00	0.95	0.14	–
	$\lambda_{y1} = 1$	0.00	0.96	0.81	1.00	–	–	–	1.00
	$\lambda_{y2} = 1$	0.00	0.96	0.82	1.00	–	–	–	1.00
2	$\sigma_e^2 = 1$	0.00	0.94	0.28	–	0.34	0.00	0.24	–
	$\psi_1 = 0.4$	0.00	0.96	0.20	–	–	–	–	–
	$\psi_2 = 0.4$	0.00	0.94	0.20	–	–	–	–	–
	$\gamma_y = 1$	0.00	0.95	0.17	–	0.00	0.93	0.15	–
	$\lambda_{y1} = 1$	0.00	0.97	0.41	1.00	–	–	–	1.00
	$\lambda_{y2} = 1$	0.00	0.94	0.41	1.00	–	–	–	1.00
3	$\sigma_e^2 = 1$	0.00	0.93	0.27	–	0.37	0.00	0.24	–
	$\psi_1 = 0.6$	0.00	0.94	0.18	–	–	–	–	–
	$\psi_2 = 0.6$	0.00	0.94	0.17	–	–	–	–	–
	$\gamma_y = 1$	0.00	0.95	0.18	–	0.00	0.94	0.15	–
	$\lambda_{y1} = 1$	0.00	0.97	0.27	1.00	–	–	–	1.00
	$\lambda_{y2} = 1$	0.00	0.93	0.27	1.00	–	–	–	1.00
4	$\sigma_e^2 = 1$	0.00	0.94	0.61	–	1.37	0.00	0.42	–
	$\psi_1 = 0.4$	0.00	0.95	0.19	–	–	–	–	–
	$\psi_2 = 0.4$	0.00	0.93	0.19	–	–	–	–	–
	$\gamma_y = 1$	0.00	0.96	0.25	–	0.00	0.94	0.19	–
	$\lambda_{y1} = 2$	−0.01	0.96	0.67	1.00	–	–	–	1.00
	$\lambda_{y2} = 2$	−0.01	0.94	0.66	1.00	–	–	–	1.00
5	$\sigma_e^2 = 1$	0.00	0.95	0.41	–	0.72	0.00	0.31	–
	$\psi_1 = 0.4$	0.00	0.95	0.22	–	–	–	–	–
	$\psi_2 = 0.4$	0.00	0.94	0.18	–	–	–	–	–
	$\gamma_y = 1$	0.00	0.95	0.20	–	0.00	0.95	0.16	–
	$\lambda_{y1} = 0.5$	0.00	0.97	0.41	1.00	–	–	–	1.00
	$\lambda_{y2} = 2$	−0.01	0.93	0.57	1.00	–	–	–	1.00

$\nu_{xs} = \nu_{xt} = 0$, $\lambda_{x2} = \lambda_{x3} = \lambda_{x4} = \lambda_{x6} = \lambda_{x7} = \lambda_{x8} = 1$, $\nu_y = 0$, and define a threshold in `Mplus` such that $P(x_{is} = 0) = P(x_{it} = 0) = 0.5$. Finally, the values of λ_{y1}, λ_{y2}, γ_y, σ_e^2, ψ_1, and ψ_2 are assumed to vary as described in Table 2.

The results of fitting the SEM and MEM to these data are given in Table 2. The results are similar to those we saw with a single latent variable, with more extreme biases associated with σ_e^2 using the MEM approach. The type-1 error rates in the SEM setting are 0.05 and 0.06 for λ_{y1} and λ_{y2}, respectively, while the type-1 error rates in the MEM setting are 0.04 and 0.05 for σ_{b1}^2 and σ_{b2}^2, respectively.

Finally, we simulate data according to an SEM with two correlated latent variables, where the model is the same as described above for two uncorrelated latent variables, with the exception that it is assumed $\text{cov}(\zeta_{1i}, \zeta_{2i}) = \psi_{12}$. Data are simulated under two models specified by $\psi_{12} = 0.1$ and $\psi_{12} = 0.2$, where in both cases $\psi_1 = \psi_2 = 0.4$. These models correspond to correlations between

Table 3 Simulation results under SEM with two correlated latent variables.

Data	True value	SEMs				MEMs			
		Bias	CR	CI	Power	Bias	CR	CI	Power
1	$\sigma_e^2 = 1$	0.00	0.93	0.28	–	0.37	0.00	0.24	–
	$\psi_1 = 0.4$	0.00	0.95	0.20	–	–	–	–	–
	$\psi_2 = 0.4$	0.00	0.94	0.20	–	–	–	–	–
	$\gamma_y = 1$	0.00	0.95	0.18	–	0.00	0.93	0.15	–
	$\lambda_{y1} = 1$	0.00	0.94	0.42	1.00	–	–	–	1.00
	$\lambda_{y2} = 1$	−0.01	0.93	0.42	1.00	–	–	–	1.00
	$\psi_{12} = 0.1$	0.00	0.96	0.09	–	–	–	–	–
2	$\sigma_e^2 = 1$	0.00	0.93	0.28	–	0.40	0.00	0.25	–
	$\psi_1 = 0.4$	0.00	0.94	0.20	–	–	–	–	–
	$\psi_2 = 0.4$	0.00	0.94	0.20	–	–	–	–	–
	$\gamma_y = 1$	0.00	0.95	0.18	–	0.00	0.94	0.15	–
	$\lambda_{y1} = 1$	0.00	0.95	0.49	1.00	–	–	–	1.00
	$\lambda_{y2} = 1$	0.00	0.95	0.49	1.00	–	–	–	1.00
	$\psi_{12} = 0.2$	0.00	0.95	0.10	–	–	–	–	–

Table 4 Simulation results under MEM with single-level clustering.

Data	True value	SEMs				MEMs			
		Bias	CR	CI	Power	Bias	CR	CI	Power
1	$\sigma_e^2 = 1$	0.12	0.42	0.21	–	0.00	0.96	0.18	–
	$\sigma_b^2 = 0.2$	–	–	–	0.65	−0.01	0.94	0.33	1.00
	$\gamma = 1$	0.00	0.93	0.13	–	0.00	0.94	0.13	–
2	$\sigma_e^2 = 1$	0.22	0.23	0.25	–	0.00	0.94	0.18	–
	$\sigma_b^2 = 0.4$	–	–	–	0.76	0.00	0.95	0.65	1.00
	$\gamma = 1$	0.00	0.94	0.15	–	0.00	0.95	0.13	–
3	$\sigma_e^2 = 1$	0.33	0.17	0.28	–	0.00	0.94	0.18	–
	$\sigma_b^2 = 0.6$	–	–	–	0.77	−0.02	0.95	0.94	1.00
	$\gamma = 1$	0.00	0.95	0.15	–	0.00	0.94	0.12	–

latent variables of 0.25 and 0.5, respectively. The results are reported in Table 3. Again power is high under both the SEM and MEM, and similar bias is observed under the MEM for σ_e^2. The corresponding type-1 error rates are 0.05 and 0.06 for both SEM parameters, λ_{y1} and λ_{y2}, and 0.04 and 0.05 for the two MEM variance parameters.

Next we simulate data under an MEM with a single clustering variable, as described by Eq. (4). We again assume $S = 4$ SNPs, each coded as binary indicators with $P(x_{is} = 1) = 0.5$ and minimal correlation induced by the assumption $P(x_{is+1} = 1|x_{is} = 1) = 0.6$ and $P(x_{is+1} = 1|x_{is} = 0) = 0.4$ for $s = 1,2,3$. A single continuous covariate $z_i \sim N(0, 1)$ is generated and we set $\nu = 0$. The remaining parameters, σ_e^2, σ_b^2 and γ are varied across the simulations as given in Table 4. The results of fitting SEMs and MEMs to these data are reported in Table 4. In this setting, a test of $\lambda_y = 0$ has reduced power for detecting association. The type-1 error rates are 0.05 and 0.03 for the SEM and MEM approaches, respectively.

Table 5 Simulation results under MEM with two-level clustering.

Data	True value	SEMs				MEMs			
		Bias	CR	CI	Power	Bias	CR	CI	Power
1	$\sigma_e^2 = 1$	0.31	0.03	0.23	–	0.00	0.95	0.18	–
	$\sigma_{b1}^2 = 0.2$	–	–	–	0.65	−0.01	0.98	0.45	0.97
	$\sigma_{b2}^2 = 0.2$	–	–	–	0.68	−0.01	0.96	0.44	0.97
	$\gamma_y = 1$	0.00	0.94	0.15	–	−0.01	0.94	0.25	–
2	$\sigma_e^2 = 1$	0.59	0.00	0.29	–	0.00	0.94	0.18	–
	$\sigma_{b1}^2 = 0.4$	–	–	–	0.78	−0.01	0.95	0.79	0.99
	$\sigma_{b2}^2 = 0.4$	–	–	–	0.79	−0.02	0.96	0.77	1.00
	$\gamma = 1$	0.00	0.94	0.16	–	0.00	0.94	0.25	–
3	$\sigma_e^2 = 1$	0.95	0.00	0.36	–	0.00	0.95	0.18	–
	$\sigma_{b1}^2 = 0.6$	–	–	–	0.77	−0.03	0.95	1.10	1.00
	$\sigma_{b2}^2 = 0.6$	–	–	–	0.77	0.00	0.95	1.11	0.99
	$\gamma_y = 1$	0.00	0.95	0.18	–	0.00	0.94	0.25	–
4	$\sigma_e^2 = 1$	0.47	0.00	0.27	–	0.00	0.97	0.18	–
	$\sigma_{b1}^2 = 0.2$	–	–	–	0.73	−0.01	0.97	0.48	0.97
	$\sigma_{b2}^2 = 0.4$	–	–	–	0.74	−0.02	0.96	0.73	1.00
	$\gamma_y = 1$	0.00	0.96	0.15	–	0.00	0.96	0.25	–

Finally, we generate data according to a two-level clustering MEM, as described by Eq. (5). Here we again assume that we observe 4 SNPs in each of 2 genes. The corresponding results of fitting SEMs and MEMs to these data are given in Table 5. Again the power for the SEM approach to detect association based on a test of λ_y is relatively small in all cases while the power for the MEM approach is reasonable (>90%) for $\sigma_b^2/\sigma_e^2 \geq 0.16$. The estimated type-1 error rates are 0.03 and 0.05 for the SEM parameters and 0.01 for both of the MEM parameters.

3.2 Genetics of therapy-associated lipid abnormalities in HIV

In this section we apply the SEM and MEM frameworks to data arising from the New Works Concept Sheet (NWCS) 224 study, an investigation of genetic factors that contribute to anti-retroviral-associated dyslipidemia in HIV-1-infected individuals. This cross-sectional study is comprised of $n = 626$ HIV-infected participants enrolled in 5 AIDS Clinical Trials Group (ACTG) trials who agreed to genetic testing. A complete discussion of the study design and patient demographics is given in Foulkes et al. (2006). Here we focus on 7 SNPs – rs1045642, rs2032582, rs2235035, rs11772987, rs10256836, rs9282564 and rs2157926 – within the ATP-binding cassette, sub-family B (MDR/TAP), member 1 (ABCB1) gene, a gene involved in transporting substrates, including Protease Inhibitors (PIs) across the cell membrane, and an additional 3 SNPs – rs2854117, rs4520 and rs2070669 – in apolipoprotein C-III (APOC3), a gene involved in inhibiting hepatic uptake of triglyceride-rich particles. All SNPs are treated as binary indicators for the presence of at least one variant allele. Interest is in characterizing association between these SNPs and high-density lipoprotein cholesterol (HDL-C). White/non-Hispanic, Hispanic and Black/non-Hispanic subjects ($n = 532$) with complete data, including known drug exposure histories, are included in analysis.

The results of fitting unadjusted models are reported in Table 6. Here we consider three models: two single-gene models (that include either ABCB1 or APOC3) and one two-gene model (that

Table 6 Application to study of therapy-associated lipid abnormalities in HIV.

	ABCB1 – model		APOC3 – model		(ABCB1, APOC3) – model	
	SEM Est (p-value)	MEM Est (p-value)	SEM Est (p-value)	MEM Est (p-value)	SEM Est (p-value)	MEM Est (p-value)
λ_y	−0.06 (0.002)	–	0.02 (0.53)	–	−0.06 (0.003); 0.01 (0.84)	–
ψ	0.78	–	0.18	–	0.78, 0.28	–
σ_b^2	–	0.004 (0.02)	–	0.00 (0.50)	–	0.004 (0.02); 0.00 (0.50)
σ_e^2	0.097	0.096	0.099	0.099	0.097	0.096

includes both ABCB1 and APOC3). The SEM and MEM results are consistent with one another, suggesting that ABCB1 is associated with HDL-C, as measured by $\widehat{\lambda}_y = -0.06$ ($p = 0.002$ and $p = 0.003$) in the SEM and $\widehat{\sigma}_b = 0.004$ ($p = 0.02$) in the MEM for both the single-gene and two-gene models. These effects are attenuated (and no longer statistically significant) in adjusted models and may represent spurious associations resulting from population-admixture, i.e. confounding by race/ethnicity and study site. Adjusted models also included PI exposure as a three-level factor – no current PI exposure; currently exposed to a non-RTV-containing PI regimen; and currently exposed to an RTV-containing PI regimen – gender, race/ethnicity and study.

To further explore the results of this model fitting procedure, we consider $\widehat{\Lambda}_x$ from the SEM and the empirical Bayes estimates of the random effects from the MEM. For illustration, we focus on the unadjusted model with the single-gene ABCB1. Based on the SEM, the relationship between the SNPs and the latent gene variable, u_i, is estimated by $\widehat{\Lambda}_x = (1.00, 1.15, -0.07, -0.50, 0.04, 0.85, -0.54)$ corresponding to rs1045642, rs2032582, rs2235035, rs11772987, rs10256836, rs9282564 and rs2157926, respectively. Associated p-values based on a Wald test are given by $p = (-, 0.00, 0.329, 0.00, 0.569, 0.00, 0.00)$ where p-values of "0" are less than 1×10^{-8}. Recall the first element of Λ is fixed at 1 for identifiability. These results suggest variant alleles at rs2032582 ($p < 1 \times 10^{-8}$) and rs9282564 ($p < 1 \times 10^{-8}$) are significantly positively associated with u_i, while variant alleles at rs11772987 ($p < 1 \times 10^{-8}$) and rs2157926 ($p < 1 \times 10^{-8}$) are significantly negatively associated with u_i. Further, $\widehat{\lambda}_y = -0.06$, suggesting an inverse relationship between u_i and HDL-C.

In total there are 46 observed genotype groups and the corresponding empirical Bayes estimates based on the MEM range from −0.097 to 0.074, as illustrated in Figure 3. All corresponding 95% prediction intervals for the genotype groups cover zero, with the exception of the group with at least one variant allele at each of the three SNPs, rs1045642, rs2032582 and rs11772987 and homozygous for the common allele at all other SNPs. Membership to this group is inversely associated with HDL-C with a corresponding empirical Bayes estimate of −0.097 and a 95% prediction interval of (−0.185,−0.009).

3.3 An alternative Bayesian network analysis framework

BN analysis is an alternative analysis framework that similarly aims to identify and characterize association among combinations of potential predictor variables and an observed trait. In this section, we briefly describe the application of one such approach, proposed by Malovini et al. (2009). This strategy is comprised of two stages: First, meta-variables are created based on fitting a

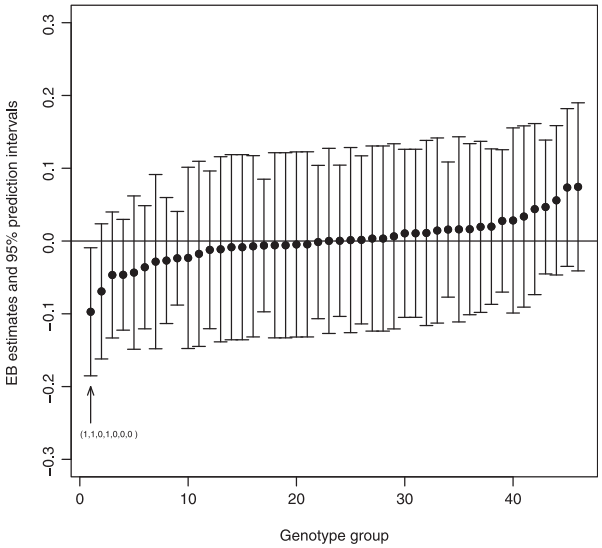


Figure 3 Empirical Bayes (EB) estimates of latent genotype group effects.

classification or regression tree to the trait under study where SNPs and other potentially relevant variables are treated as potential predictor variables; and second, a BN is learned based on these meta-variables and the trait under study. Analysis is based on application of the `rpart()` and `network()` functions within the `R` `rpart` and `deal` packages, respectively.

The results of applying the approach of Malovini et al. (2009), to the SNPs within ABCB1 and APOC3 separately, and in combination, are provided in Fig. 4. To begin, we fitted a regression tree to the log transformed quantitative trait, HDL-C separately for each gene. For ABCB1, a single split is observed, as illustrated in Fig. 4A, where ABCB1.S7 corresponds to rs2157926, and a cost complexity parameter (cp) of 0.01 is applied for first-stage pruning. This constitutes the meta-variable used for the second-stage BN analysis of ABCB1. For APOC3, no splits result in an increase of more than cp = 0.01 in the overall R-squared value, and thus individual SNP variables for this gene are used in the second-stage BN analysis. The resulting directed acyclic graphs (DAGs) illustrated in Fig. 4B–D are consistent with the results presented in Table 6 based on the SEM and MEM analyses. Again, an association between ABCB1 and HDL-C is observed (Fig. 4B), while associations between SNPs within APOC3 and HDL-C are not detectable (Fig. 4C). The DAG based on both genes (Fig. 4D) additionally identifies an association between ABCB1 and the first SNP within APOC3.

Finally, as a case study, we applied the two-stage BN approach to a single randomly selected simulated data set from each of the scenarios (i.e. combinations of parameters) described in Tables 1–5. The results are presented in Fig. 5. In-line with the finding presented in Table 1 of consistently high power of the SEM and MEM under the SEM model with a single latent variable, Fig. 5A illustrates that all five corresponding DAGs identify association between the gene metavariable and the trait. Under the SEMs with two uncorrelated latent variables, the BN analysis consistently identifies at least one of the two gene metavariables; however, in 3 of the 5 cases only one gene is identified, as shown in Fig. 5B. The association between the two genes under the SEM with correlated latent variables is detected with higher correlation, as illustrated in Fig. 5C. Finally, the BN analysis appropriately identifies the gene metavariable for the MEM with single-level clustering, and, in 3 of the 4 cases, was able to identify both gene metavariables under the MEM with two-level clustering.

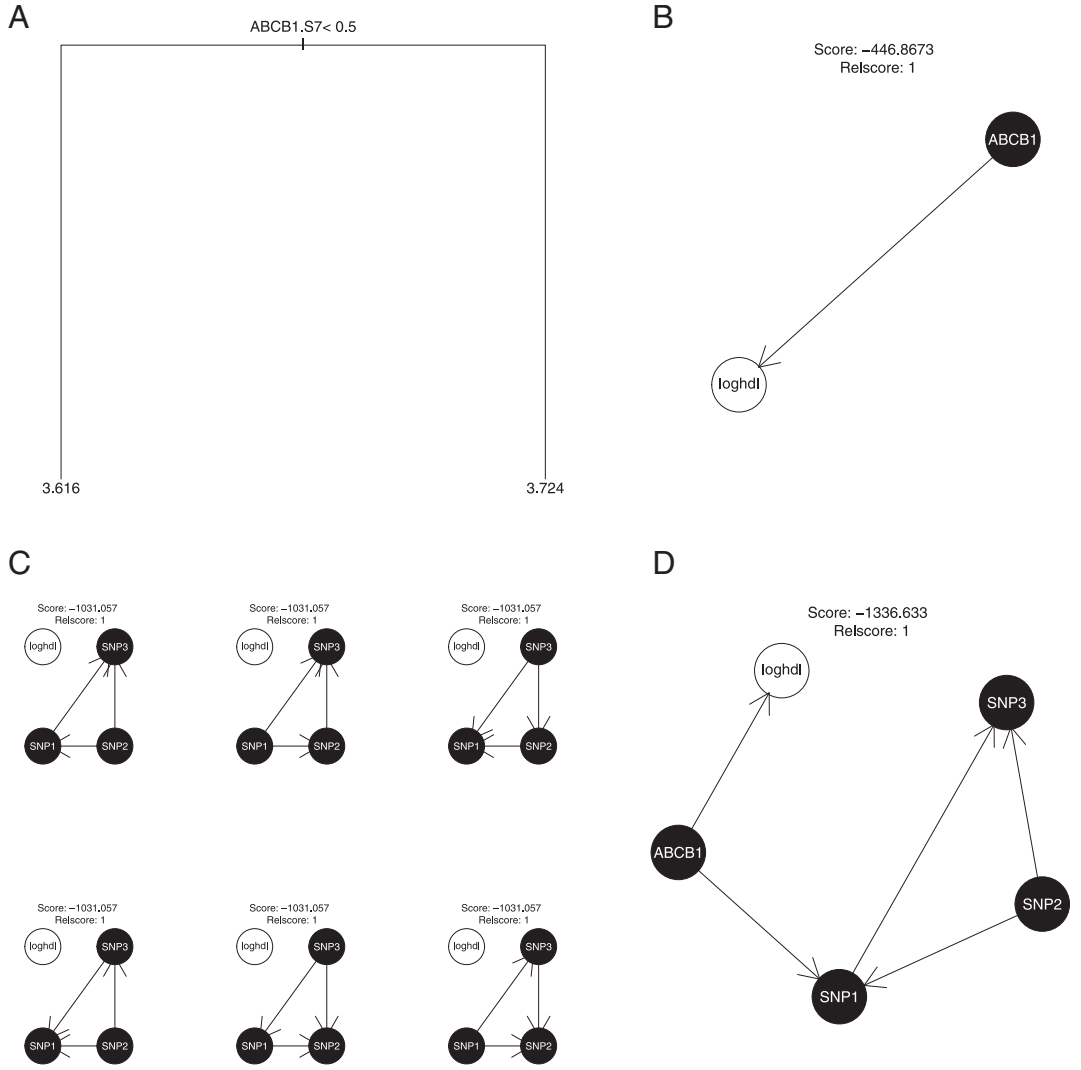


Figure 4 Bayesian networks of ABCB1, APOC3 and HDL-C. (A) Fitted regression tree using all SNPs within ABCB1 (coded as binary indicators for the presence of at least one variant allele) as potential predictor variables and log-transformed HDL-C as the outcome. (B) DAG based on ABCB1 metavariable and log-transformed HDLC. (C) DAGs with highest scores based on individual SNPs within APOC3 and log-transformed HDL-C. (D) DAG based on ABCB1 metavariable, individuals SNPs within APOC3 and log-transformed HDL-C.

4 Discussion

In this manuscript we describe the application of two related latent variable modeling approaches, MEMs and SEMs, for identifying and characterizing genetic contributors to complex diseases. While these two frameworks have some important commonalities, several notable differences emerged during our investigation. These are highlighted by the assumptions listed in Section 2.3,

Figure 5 Bayesian networks of simulated data. (A) DAGs based on data simulated according to a SEM with one latent variable where gene represents a metavariable based on a fitted regression tree. The 5 DAGs correspond to each combination of parameters given in Table 1. (B) DAGs based on data simulated according to a SEM with two uncorrelated latent variables where gene1 and gene2 represent metavariables based on two separate fitted regression trees. The 5 DAGs correspond to each combination of parameters given in Table 2. (C) DAGs based on data simulated according to a SEM with two correlated latent variables where gene1 and gene2 represent metavariables based on two separate fitted regression trees. The 2 DAGs correspond to each combination of parameters given in Table 3. (D) DAGs based on data simulated according to an MEM with single-level clustering where gene represents a metavariable based on a fitted regression tree. The 3 DAGs correspond to each combination of parameters given in Table 4. (E) DAGs based on data simulated according to an MEM with two-level clustering where gene1 and gene2 represent metavariables based on two separate fitted regression trees. The 4 DAGs correspond to each combination of parameters given in Table 5.

under which the SEM reduces to the MEM as described for this setting. Importantly, in the context of an SEM, the test of association is based on a fixed effect coefficient (λ_y) relating the latent gene variable to the trait. In the MEM context, on the other hand, the test of association is based on a variance parameter (σ_b^2) of the latent gene effects.

Interestingly, our simulation studies reveal that the performance of the two modeling approaches is comparable under the SEMs in terms of power and type-1 error rates; however, when the data arise from an MEM, power for the SEM approach is lower than the corresponding power for detecting association using the MEM approach. Also of note, when data are simulated under the SEM model, the estimates of the nuisance parameter, σ_e, under the MEM exhibit substantial bias. In these cases, the estimate of σ_b is also estimated poorly (results not shown). Notably, for Tables 1 and 2, $\text{var}(y) = \lambda_y^2\psi + \sigma_e^2$ and an estimate of this variance under the MEM is $\widehat{\sigma}_e^2 + \widehat{\sigma}_b^2$. For example, for the first scenario in Table 1, we have $\text{var}(y) = \lambda_y^2\psi + \sigma_e^2 = 1 \times 0.2 + 1 = 1.2$ and the estimate under the MEM is $\widehat{\text{var}}(y) = \widehat{\sigma}_e^2 + \widehat{\sigma}_b^2 = 1.13 + 0.06 = 1.19$ (results not shown). In general, these are not as closely aligned; however, there appears to be a tradeoff between the two parameters. In turn, estimation of γ_y depends on $\widehat{\text{var}}(y)$. Further research on sensitivity to alternative underlying model specifications may further elucidate the relative merits of each approach. Specifically, SEMs may be more conducive to testing specific hypotheses involving multiple genes and their relationships to one another in a pathway to disease.

A notable limitation of both the MEM and SEM approaches is their potential inability to handle a large number of SNPs. In the context of the MEM, the number of genotype groups can become unwieldy as the number of SNPs increases, as described by Foulkes et al. (2008). Interestingly, inclusion of highly correlated SNPs in the MEM approach results in fewer genotype groups but does not otherwise effect model performance. Furthermore, our preliminary research suggests that iteratively sampling a subset of SNPs and fitting the MEM, and then combining results over the multiple samples (Efron, 1979, 1981; Good, 2005), performs reasonably well (results not shown) in terms of the overall power for detecting association in the setting of a large number of SNPs. The extension of the MEM approach to more than one clustering variable, as described in Section 2.2, also offers the advantage of reducing the total number of genotype groups. For example, if we observe r SNPs in one gene and s SNPs in a second gene, the original formulation of the MEM approach involves $2^{(r+s)}$ genetic groups while the proposed extension involves only $(2^r + 2^s)$ genetic groups (across two clustering variables). Additional research is needed to evaluate the performance of this extended MEM approach with multiple SNPs across a larger number of genes.

We also presented the results of applying an alternative two-stage BN analysis approach to the NWCS224 data, as well as to randomly selected simulated data sets. In the real data example, we were unable to fit a regression tree with splits beyond the root node subject to the specified pruning criterion for one of the genes under study. In this case, the analysis reduced to fitting a BN to single SNP variables rather than metavariables as described by Malovini et al. (2009). On the other hand, for the ABCB1 gene, a metavariable did emerge, albeit based on a single SNP, and an association was detected. Importantly, the structure of association that CART is designed to detect, namely conditional associations (Foulkes, 2009), may explain why only a single SNP emerged while for the SEM four SNPs within this gene were identified as statistically relevant. Although our case studies of simulated data suggest that the BN analysis generally (though not always consistently) identified relevant genes, further investigation is needed to characterize the overall performance and type-1 error rates.

Finally, further research is needed to characterize the performance of both the SEM and MEM frameworks in the context of more complex structures of association and underlying genetic models. In a recent manuscript, we describe application of a mixture modeling approach that may be more appropriate than the single normal prior assumption on the random effects under a dominant or recessive genetic model (Au et al., 2010). While the inclusion of cross-classified clusters, as proposed in Section 2.2, allows for consideration of more SNPs through reduction in the number of genotype

groups, this approach assumes an additive model of association. More generally, it is of interest to characterize interactions among genes in their effect on the trait under study, an area of on-going research.

Acknowledgements The support for this research was provided by National Institutes of Health (NIH) individual research awards, NIH/NHLBI R01HL107196 and NIH/NIDDK R01DK021224.

Conflict of interest
The authors have declared no conflict of interest.

References

Au, K., Lin, R. and Foulkes, A. S. (2011). Mixture modeling as an exploratory framework for genotype-trait associations. *Journal of the Royal Statistical Society Series C* **60**, 355–375.

Bastone, L., Reilly, M., Rader, D. and Foulkes, A. (2004). MDR and PRP: a comparison of methods for high-order genotype–phenotype associations. *Human Heredity* **58**, 82–92.

Bentler, P. M. and Weeks, D. G. (1980). Linear structural equations with latent variables. *Psychometrika* **45**, 289–308.

Breiman, L. (2001). Random forests. *Machine Learning* **45**, 5–32.

Breiman, L., Friedman, J., Olshen, R. and Stone, C. (1993). *Classification and Regression Trees*. Chapman and Hall/CRC, New York, USA.

Bureau, A., Dupuis, J., Falls, K., Lunetta, K., Hayward, B., Keith, T. and Van Eerdewegh, P. (2005). Identifying SNPs predictive of phenotype using random forests. *Genetic Epidemiology* **28**, 171–182.

Efron, B. (1979). Bootstrap methods: another look at the jackknife. *Annals of Statistics* **7**, 1–26.

Efron, B. (1981). Nonparametric estimates of standard error: the jackknife, the bootstrap and other methods. *Biometrika* **68**, 589–599.

Foulkes, A. (2009). *Applied Statistical Genetics with R: for Population-based Association Studies*. Springer, NY.

Foulkes, A. S. and De Gruttola, V. (2002). Characterizing the relationship between HIV-1 genotype and phenotype: prediction-based classification. *Biometrics* **58**, 145–156.

Foulkes, A. S., De Gruttola, V. and Hertogs, K. (2004). Combining genotype groups and recursive partitioning: an application to human immunodeficiency virus type 1 genetics data. *Journal of the Royal Statistical Society Series C* **53**, 311–323.

Foulkes, A. S., Reilly, M., Zhou, L., Wolfe, M. and Rader, D. J. (2005). Mixed modelling to characterize genotype–phenotype associations. *Statistics in Medicine* **24**, 775–789.

Foulkes, A. S., Wohl, D. A., Frank, I., Puleo, E., Restine, S., Wolfe, M. L., Dube, M. P., Tebas, P. and Reilly, M. P. (2006). Associations among race/ethnicity, ApoC-III genotypes, and lipids in HIV-1-infected individuals on antiretroviral therapy. *PLoS Medicine* **3**, 0337–0347.

Foulkes, A. S., Yucel, R. and Li, X. (2008). A likelihood-based approach to mixed modeling with ambiguity in cluster identifiers. *Biostatistics* **9**, 635–657.

Gentleman, R., Carey, V., Huber, W., Irizarry, R. and Dudoit, S. (2005). *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*. Springer, New York, USA.

Goeman, J., van de Geer, S., de Kort, F. and van Houwelingen, H. (2004). A global test for groups of genes: testing association with a clinical outcome. *Bioinformatics* **20**, 93–99.

Good, P. I. (2005). *Permutation, Parametric, and Bootstrap Tests of Hypotheses*. Springer, New York, USA.

Hartigan, J. A. (1975). *Clustering Algorithms*. Wiley, Ann Arbor.

Hastie, T., Tibshirani, R. and Friedman, J. (2001). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, New York, USA.

Jöreskog, K. G. (1973). A general method for estimating a linear structural equation system. In: *Structural Equation Models in the Social Sciences* (ed. A. S. Goldberger and O. D. Duncan), New York, pp. 85–112.

Kooperberg, C., LeBlanc, M. and Obenchain, V. (2010). Risk prediction using genome-wide association studies. *Genetic Epidemiology* **34**, 643–52.

Kooperberg, C. and Ruczinski, I. (2005). Identifying interacting SNPs using Monte Carlo logic regression. *Genetic Epidemiology* **28**, 157–170.

Kooperberg, C., Ruczinski, I., LeBlanc, M. and Hsu, L. (2001). Sequence analysis using logic regression. *Genetic Epidemiology* **21**, S626–S631.

Laird, N. and Ware, J. (1982). Random-effects models for longitudinal data. *Biometrics* **38**, 963–974.

Lee, S., Jhun, M., Lee, E. K. and Park, T. (2007). Application of structural equation models to construct genetic networks using differentially expressed genes and single-nucleotide polymorphisms. *BMC Proceedings* **1**, S76.

Lee, S. and Shi, J. (2001). Maximum likelihood estimation of two-level latent variable models with mixed continuous and polytomous data. *Biometrics* **57**, 787–794.

Malovini, A., Nuzzo, A., Ferrazzi, F., Puca, A. and Bellazzi, R. (2009). Phenotype forecasting with SNPs data through gene-based Bayesian networks. *BMC Bioinformatics* **10**, S7.

Muthén, B. O. (1984). A general structural equation model with dichotomous, ordered categorical and continuous latent indicators. *Psychometrika* **49**, 115–132.

Muthén, B. O. (2002). Beyond SEM: general latent variable modeling. *Behaviormetrika*, **29**, 81–117.

Muthén, L. K. and Muthén, B. O. (2007). *Mplus User's Guide*. Muthén & Muthén, Los Angeles, CA.

Nock, N. L., Larkin, E. K., Morris, N. J., Li, Y. and Stein, C. M. (2007). Modeling the complex gene x environment interplay in the simulated rheumatoid arthritis GAW15 data using latent variable structural equation modeling. *BMC Proceedings* **1**, S118.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufmann, San Mateo, CA.

Pugesek, B. H., Tomer, A. and Eye, A. V. (2003). *Structural Equation Modeling: Applications in Ecological and Evolutionary Biology*. Cambridge University Press, Cambridge.

Rabe-Hesketh, S., Skrondal, A. and Pickles, A. (2004). Generalized multilevel structural equation modeling. *Psychometrika* **69**, 167–190.

Reboussin, B. and Liang, K. (1998). An estimating equations approach for the liscomp model. *Psychometrika* **63**, 165–182.

Rodin, A. and Boerwinkle, E. (2005). Mining genetic epidemiology data with Bayesian networks i: Bayesian networks and example application (plasma apoE levels). *Bioinformatics* **21**, 3273–3278.

Ruczinski, I., Kooperberg, C. and LeBlanc, M. (2003). Logic regression. *Journal of Computational and Graphical Statistics* **12**, 475–511.

Ruczinski, I., Kooperberg, C. and LeBlanc, M. (2004). Exploring interactions in high dimensional genomic data: an overview of logic regression. *Journal of Multivariate Analysis* **90**, 178–195.

Sánchez, B. N., Budtz-Jørgensen, E., Ryan, L. M. and Hu, H. (2005). Structural equation models: a review with applications to environmental epidemiology. *Journal of the American Statistical Association* **100**, 1443–1455.

Schwender, H. and Ickstadt, K. (2008). Identification of SNP interactions using logic regression. *Biostatistics* **9**, 187–198.

Schwender, H., Ickstadt, K. and Rahnenführer, J. (2008). Classification with high-dimensional genetic data: assigning patients and genetic features to known classes. *Biometrical Journal* **50**, 911–926.

Segal, M., Barbour, J. and Grant, R. (2004). Relating HIV-1 sequence variation to replication capacity via trees and forests. *Statistical Applications in Genetics and Molecular Biology* **3**, Article 2.

Skrondal, A. and Rabe-Hesketh, S. (2004). *Generalized Latent Variable Modeling: Multilevel, Longitudinal, and Structural Equation Models*. Chapman & Hall/CRC, Boca Raton.

Skrondal, A. and Rabe-Hesketh, S. (2005). Structural equation modeling: categorical variables. In: *Encyclopedia of Statistics in Behavioral Science*, Wiley, London pp. 1–8.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B* **58**, 267–288.

Wu, T., Chen, Y., Hastie, T., Sobel, E. and Lange, K. (2009). Genome-wide association analysis by Lasso penalized logistic regression. *Bioinformatics* **25**, 714–21.

Zhang, H. and Singer, B. (1999). *Recursive Partitioning in the Health Sciences*. Springer, New York, USA.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B* **67**, 301–320.

