Biometrics 68, 466–476
June 2012
DOI: 10.1111/j.1541-0420.2011.01677.x

A Latent Variable Approach to Study Gene–Environment Interactions in the Presence of Multiple Correlated Exposures

Brisa N. Sánchez,* Shan Kang, and Bhramar Mukherjee

Department of Biostatistics, University of Michigan, Ann Arbor, Michigan 48109, U.S.A.
**email:* brisa@umich.edu

Summary. Many existing cohort studies initially designed to investigate disease risk as a function of environmental exposures have collected genomic data in recent years with the objective of testing for gene–environment interaction ($G \times E$) effects. In environmental epidemiology, interest in $G \times E$ arises primarily after a significant effect of the environmental exposure has been documented. Cohort studies often collect rich exposure data; as a result, assessing $G \times E$ effects in the presence of multiple exposure markers further increases the burden of multiple testing, an issue already present in both genetic and environment health studies. Latent variable (LV) models have been used in environmental epidemiology to reduce dimensionality of the exposure data, gain power by reducing multiplicity issues via condensing exposure data, and avoid collinearity problems due to presence of multiple correlated exposures. We extend the LV framework to characterize gene–environment interaction in presence of multiple correlated exposures and genotype categories. Further, similar to what has been done in case–control $G \times E$ studies, we use the assumption of gene–environment (*G-E*) independence to boost the power of tests for interaction. The consequences of making this assumption, or the issue of how to explicitly model *G-E* association has not been previously investigated in LV models. We postulate a hierarchy of assumptions about the LV model regarding the different forms of *G-E* dependence and show that making such assumptions may influence inferential results on the G, E, and $G \times E$ parameters. We implement a class of shrinkage estimators to data adaptively trade-off between the most restrictive to most flexible form of *G-E* dependence assumption and note that such class of compromise estimators can serve as a benchmark of model adequacy in LV models. We demonstrate the methods with an example from the Early Life Exposures in Mexico City to Neuro-Toxicants Study of lead exposure, iron metabolism genes, and birth weight.

Key words: Gene–environment independence; Principal components; Shrinkage estimation; Structural equation models.

1. Introduction

It is now clear from many lines of evidence that pure genetics or pure environmental factors play only a partial role in the etiology of most complex diseases. Instead, it is now accepted that the majority of chronic diseases likely stem from interactions between genetic traits, "*G*," and environmental factors, "*E*"—an exponentially growing area of study (Khoury and Wacholder, 2009). Characterizing gene–environment interactions, "$G \times E$ effects," is critical in understanding the biological mechanisms of disease etiology and can impact preventive medicine and public health by informing the way clinicians advise their patients and the way public health practitioners assess risk and set policy. Statistical approaches that heighten our ability to understand $G \times E$ effects can accelerate mapping of the so far elusive environmental footprint of disease etiology.

Established environmental health cohorts that have demonstrated modest health effects of the environment are now collecting genomic data to test $G \times E$ effects. However, $G \times E$ interaction studies are statistically difficult problems because of exposure measurement error, multiple potential exposure markers, and prohibitive sample sizes required to reach adequate power. Statistical methods to boost efficiency for testing $G \times E$ effects have primarily been developed for case–control studies, where imposing the assumption of independence between environmental exposures and inherited genetic susceptibility factors, so called *G-E* independence, boosts efficiency of $G \times E$ effect estimates (Chatterjee and Carroll, 2005, and references therein). Hybrid approaches that protect against bias under departures from independence constraints have also been proposed (e.g., Mukherjee and Chatterjee, 2008; Chen, Chatterjee, and Carroll, 2009; Li and Conti, 2009).

Latent variable (LV) models have been used in environmental health studies to extract features from a set of correlated biomarkers, thus reducing dimensionality of exposure data and multiple testing burden, and enhancing power (e.g., Budtz-Jørgensen et al., 2003b). These models have enormous potential for modeling gene–environment interaction between multiple genes and multiple environmental exposures, an area which is still in its formative phase. Chatterjee et al. (2006) motivate a one degree of freedom test for gene–gene interactions from an LV perspective in a case–control study. However, the issue of modeling *G-E* dependence through the LV framework is not discussed. More general LV models, as proposed here, can incorporate and estimate *G-E* association structures, which are of interest to environmental health researchers, and also impose constraints such as independence. In our motivating example, the relationship between iron metabolism genes and lead exposure is of interest in itself (e.g., Hopkins et al., 2008). Very few

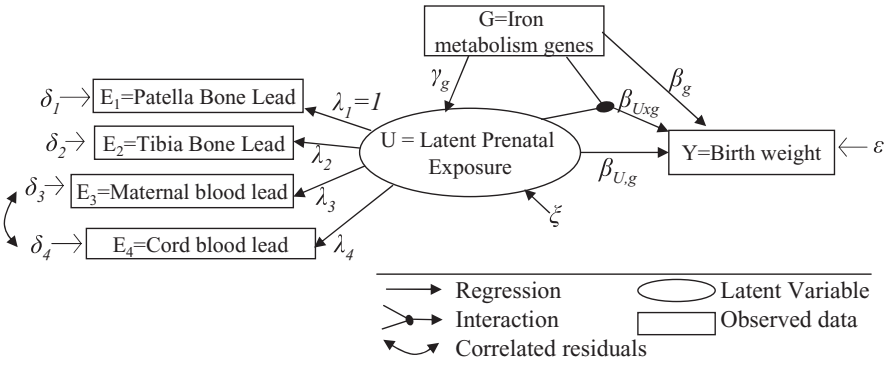

Figure 1. Path diagram showing relationships between exposure biomarkers, latent prenatal lead exposure, iron metabolism genes, and birth weight.

attempts exist that pose a general LV model for studying $G \times E$ effects in cross-sectional or cohort studies (e.g., Dhungana et al., 2007; Rathouz et al., 2008; Javaras, Hudson, and Laird, 2010), especially those investigating an array of *G-E* dependence structures.

The Early Life Exposures in Mexico City to Neurotoxicants (ELEMENT) Study motivates our work. ELEMENT consists of four longitudinal birth cohorts in Mexico City, constituting over 2000 mother infant pairs with prospectively collected exposure data and several anthropometric, cardiovascular, neurodevelopment, and behavioral outcomes. Genotyping for these cohorts is underway, with genotyping for the first cohort completed on a set of candidate genes (≈400 pairs). Figure 1 is a path diagram describing relationships between four biomarkers of prenatal lead exposure, their interaction with iron metabolism genes, and birth weight.

In Section 2, we describe a model structure to summarize a group of biomarkers into LVs, and the spectrum of *G-E* dependence models we consider. In Section 3, we describe maximum likelihood estimation (MLE) and a general class of shrinkage estimators to data adaptively compromise between the most stringent and most flexible models for *G-E* association. Small-scale simulation studies (Section 4) bring out salient features of our methodology. Section 5 presents analyses of our motivating example. Section 6 discusses the use of LV models for $G \times E$ studies, and shrinkage estimators in general LV modeling beyond the $G \times E$ context.

2. A Latent Exposure Model for $G \times E$ Studies

2.1 *Model Representation*

For the ith of N individuals, let Y_i represent a univariate health outcome (here birth weight), and $\boldsymbol{U}_i$ be an $l \times 1$ vector of latent exposures measured indirectly by a set of p measurements $\boldsymbol{E}_i$. In the example, $l = 1$, $p = 4$, and U_i is prenatal lead exposure (Figure 1). Let $\boldsymbol{Z}_i$ and $\boldsymbol{W}_i$ represent $q \times 1$ and $r \times 1$ covariate vectors, respectively. Without any loss of generality, genotype classes are represented by a categorical variable G_i with classes $g = 0, \ldots, G$. Genotype classes may arise from data on biallelic polymorphisms where a "risk" allele "A" may alter the exposure metabolism pathway or affect health (with the reference allele denoted by "a") or from combinations of genetic markers measured at multiple loci (e.g., risk alleles A, B). In the lead example we consider two single nucleotide polymorphisms (SNPs) implicated in iron metabolism, but due to sparsity of data, we assume G_i can take two possible values: zero for wild type on both SNPs ("aa" and "bb"), and 1 for at least one copy of either of the risk alleles (i.e., "Aa," "AA,", "Bb," or "BB"), consistent with dominant/recessive models for genetic susceptibility. Alternatively, genetic groups can arise as categorization of an underlying genetic risk score that combines several markers identified by existing genome-wide association studies (Qi et al., 2011), or from infant–mother genotype combinations at a single locus in studies of prenatal fetal exposure.

The LV model is then specified in two stages: a health outcome model and an exposure model. In the *outcome model*, the association between the outcome Y_i, exposure $\boldsymbol{U}_i$ and genetic category $G_i = g$ conditional on covariates $\boldsymbol{Z}_i$ is characterized by either

$$Y_i = \beta_{0,g} + \boldsymbol{\beta}_{U,g}^{\top} \boldsymbol{U}_i + \boldsymbol{\beta}_{Z,g}^{\top} \boldsymbol{Z}_i + \epsilon_i, \tag{1}$$

$$\text{or } Y_i = \beta_0 + \boldsymbol{\beta}_U^{\top} \boldsymbol{U}_i + \sum_{g=1}^{G} \left\{ \beta_g I_{gi} + \boldsymbol{\beta}_{g\times U}^{\top} I_{gi} \cdot \boldsymbol{U}_i + \boldsymbol{\beta}_Z^{\top} \boldsymbol{Z}_i + \boldsymbol{\beta}_{g\times Z}^{\top} I_{gi} \cdot \boldsymbol{Z}_i \right\} + \epsilon_i, \tag{2}$$

using genotype class indicators $I_{gi} = I(G_i = g)$. The mean-zero error ϵ_i has variance σ^2. Equation (1) is written using the multiple group notation (Bollen 1989), which is useful in writing the likelihood (Section 3), and $\beta_{0,g}, \boldsymbol{\beta}_{U,g}, \boldsymbol{\beta}_{Z,g}$ are parameters specific to class $g = 0, \ldots, G$. Such notation is standard in widely used LV software. In (1), the gene–environment interaction test is specified as H_0: $\beta_{U,0} = \cdots = \beta_{U,G}$, i.e., homogeneity of the environment's effects across genetic groups, and is equivalent to testing the interaction parameters in (2), i.e., $H_0 : \boldsymbol{\beta}_{g\times U} = \boldsymbol{\beta}_{U,g} - \boldsymbol{\beta}_{U,0} = 0$ for all $g = 1, \ldots, G$. $G \times E$ interactions are of interest in both environmental and genetic epidemiology, but in environmental epidemiology the $G \times E$ question arises after showing main effects of exposure on outcome, and heterogeneity of effects across genetic subgroups as well as the exposure effect within class g, $\boldsymbol{\beta}_{U,g}$, are of primary interest.

The exposure model consists of a *model for the latent variable* (3) dependent on covariates $\boldsymbol{W}_i$, and a *measurement model* (4) relating the observed exposure measurements to

the LV

$$\boldsymbol{U}_i = \boldsymbol{\alpha}_{0g} + \boldsymbol{\alpha}_W \boldsymbol{W}_i + \boldsymbol{\xi}_i \tag{3}$$

$$\boldsymbol{E}_i = \boldsymbol{\nu}_g + \boldsymbol{\Lambda}_g \boldsymbol{U}_i + \boldsymbol{\delta}_i \tag{4}$$

with (3) and (4) again written in the multiple group notation. Regression coefficients $\boldsymbol{\alpha}_{0g}$ and $\boldsymbol{\alpha}_W$ are $l \times 1$ and $l \times r$ matrices with $\boldsymbol{\gamma}_g = \boldsymbol{\alpha}_{0g} - \boldsymbol{\alpha}_{00}$ being effect of genotype class on exposure $\boldsymbol{U}_i$, and covariates $\boldsymbol{W}_i$ $(r \times 1)$ may help predict exposure levels for a given subject (e.g., occupation). The zero-mean error terms, $\boldsymbol{\xi}_i$, are assumed independent of ϵ_i, and have category-specific $l \times l$ covariance matrices $\boldsymbol{\Phi}_g$. Means vector $\boldsymbol{\nu}_g$ and factor loading matrix $\boldsymbol{\Lambda}_g$ are $p \times 1$ and $p \times l$, respectively, and $\boldsymbol{\delta}_i$ has zero mean and $p \times p$ covariance matrix $\boldsymbol{\Theta}_g$.

Although LV models are helpful in many respects, one well-known problem is the potential for lack of identifiability. Standard identifiability constraints have been developed for linear latent variable models (Bollen, 1989), and identifiability of latent class models has also been investigated (Huang and Bandeen-Roche, 2004). Essentially, model parameters are constrained to ensure identifiability; for example, some entries of $\boldsymbol{\nu}_g$ and $\boldsymbol{\Lambda}_g$ are fixed to 0 or 1, although sometimes algebraic proofs of identifiability are needed (Sánchez et al., 2005). In the lead example, the constraints $\boldsymbol{\nu}_g = (0, \nu_{g,2}, \nu_{g,3}, \nu_{g,4})^\top$, and $\boldsymbol{\Lambda}_g = (1, \lambda_{g,2}, \lambda_{g,3}, \lambda_{g,4})^\top$ fix the mean and scale of the latent exposure to those of patella lead. Parameters in $\boldsymbol{\Phi}_g$ are typically unconstrained, while the off-diagonal elements of $\boldsymbol{\Theta}_g$ are typically, although not necessarily, restricted to be zero denoting conditional independence between $\boldsymbol{E}_i$'s given $\boldsymbol{U}_i$. However, theoretical identifiability may not necessarily guarantee numerical stability of results, which also depends on sample sizes. Some investigators recommend at least 5 to 10 observations per parameter estimated (Westland, 2010). Hence, users need to be attentive as to what model the available sample size allows them to fit.

2.2 *Modeling G-E Dependence*

In many $G \times E$ studies, it may be natural to assume that an individual's environmental exposure is independent of genetic factors, but may not be realistic when the gene and the exposure share a metabolic pathway. For example, iron metabolism genes may increase lead absorption (Hopkins et al., 2008), and characterizing such dependence may shed insight into the mechanistic process.

Varying degrees of *G-E* dependence can be modeled through imposing further constraints on the exposure model parameters (Figure 2). The most restrictive assumption is that parameters are homogeneous across genotypes. We use "*A*0" to denote this full *G-E* independence

$$A0 : (\boldsymbol{\alpha}_{0g}, \boldsymbol{\Phi}_g, \boldsymbol{\nu}_g, \boldsymbol{\Lambda}_g, \boldsymbol{\Theta}_g) = (\boldsymbol{\alpha}_0, \boldsymbol{\Phi}, \boldsymbol{\nu}, \boldsymbol{\Lambda}, \boldsymbol{\Theta}).$$

With *A*0, the exposure model (3)–(4) has at least l parameters in each of $\boldsymbol{\alpha}_0$ and $\boldsymbol{\Phi}$, $p - l$ factor loadings $\boldsymbol{\Lambda}$ and $p - l$ intercepts $\boldsymbol{\nu}$, and p parameters $\boldsymbol{\Theta}$, for a total of at least $3p$ parameters. In the lead exposure model, *A*0 totals 13 exposure model parameters.

A first step at relaxing *A*0 is to allow the intercepts and variances in the LV equation (3) to differ by genotype. Letting $\gamma_g = \alpha_{0g} - \alpha_{00}$ be the gene effect on the latent exposure we have

$$A1 : (\boldsymbol{\nu}_g, \boldsymbol{\Lambda}_g, \boldsymbol{\Theta}_g) = (\nu, \boldsymbol{\Lambda}, \boldsymbol{\Theta}),$$
$$\text{but } \gamma_g \neq 0 \text{ or } \boldsymbol{\Phi}_g^{-1}\boldsymbol{\Phi}_0 \neq \boldsymbol{I} \text{ for at least one } g.$$

Relaxing these constraints is very natural, because genotype status may increase absorption of pollutants from the environment as well as change exposure variability. Genotype status may change the distribution of the observed exposure measures, $\boldsymbol{E}_i$, but modeling change in the distribution of the

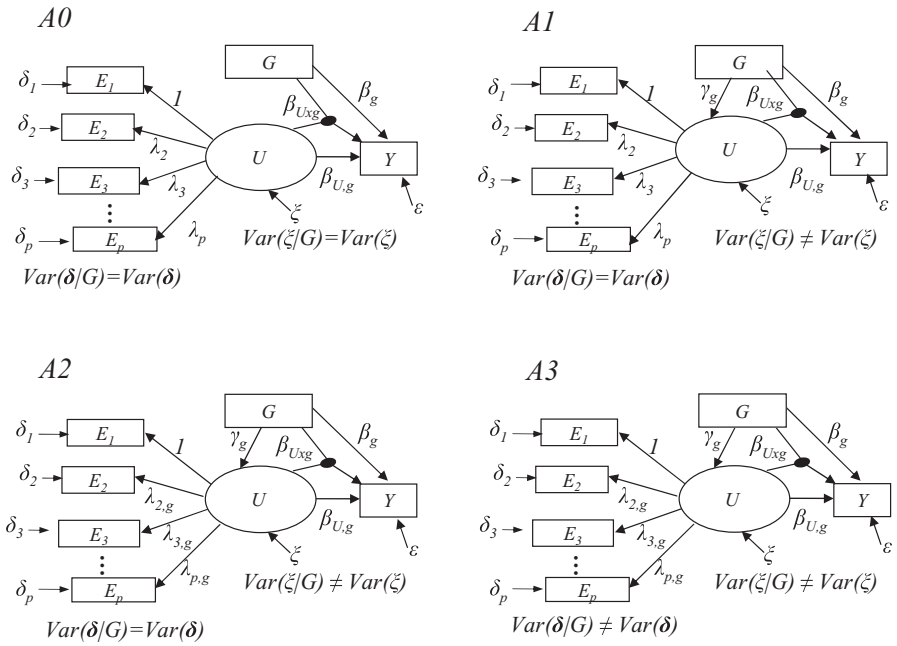

Figure 2. Path diagrams showing gene–environment dependence assumptions.

underlying exposure is a parsimonious way of modeling changes in the actual $\boldsymbol{E}_i$. This assumption has at least $3p + 2lG$ exposure model parameters; 15 in the lead example. Alternatively, one could restrict the latent variable variances $\boldsymbol{\Phi}_g$ to be equal across genotype subgroups; that is, a slightly modified assumption $A1^*: (\boldsymbol{\Phi}_g, \boldsymbol{\nu}_g, \boldsymbol{\Lambda}_g, \boldsymbol{\Theta}_g) = (\boldsymbol{\Phi}, \boldsymbol{\nu}, \boldsymbol{\Lambda}, \boldsymbol{\Theta})$.

Next, constraints of equal means $\boldsymbol{\nu}$ and factor loadings $\boldsymbol{\Lambda}$ can be removed,

$$A2: \boldsymbol{\Theta}_g = \boldsymbol{\Theta}.$$

Biological mechanisms that modify transfer of pollutants from one compartment to another are consistent with this assumption. In the lead example, three of the observed prenatal lead exposure biomarkers are measured on the mother, while umbilical cord blood on the offspring. The transfer rates from maternal compartments (i.e., blood and bone) to offspring compartment may vary by child's genotype, such that factor loading λ_4 may vary by genotype. $A2$ has at least $3p + 2Gp$ exposure model parameters; 21 in the lead example.

Lastly, all equality constraints on the parameters can be removed, namely,

$$A3: \text{ All model parameters differ by genotype,}$$

yielding at least $4p + 2Gp$ exposure model parameters; 26 in the lead example. In classical multiple group analyses (Bollen, 1989), one might additionally posit that the structure of the whole model might differ by genotype; e.g., that the number of latent variables differs by genotype. We restrict our attention to assumptions $A0$–$A3$, where the model structure is the same across genotypes.

Assumptions $A0$–$A3$ have different number of parameters; an increase from 13 to 26 parameters in the example. Constraints on exposure model parameters may increase efficiency and power for the main hypotheses (G, E or $G \times E$ tests), but could induce bias in parameters of interest if they are incorrectly assumed. Model evaluation strategies are available that may help select which assumption $A0$–$A3$ fits the data best (Bentler and Hu, 1995). However, this task may not be straightforward because many fit criteria exist for LV models. Furthermore, testing all parameter constraints may incur a high type 1 error rate, because the study will unlikely be powered to detect significant differences across genotypes in all model parameters.

3. Parameter Estimation

Although various estimation procedures have been proposed for LV models (Bollen, 1989), full MLE is the most common estimation procedure given its wide availability in software packages. We review MLE and describe the implementation of shrinkage estimators that combine MLE estimates. Using shrinkage estimators may be a more suitable approach for parameter estimation. The strategy would be to fit the most restrictive model $A0$ and the most flexible model $A3$ (or $A2$ depending on sample size), and then use shrinkage to derive the final composite estimates.

3.1 *Maximum Likelihood Estimation*

Let $\boldsymbol{Y}_i^* = (Y_i, \boldsymbol{E}_i^\top)^\top$ and $\boldsymbol{\theta}$ be the vector of all model parameters. Assuming that ϵ_i, $\boldsymbol{\xi}_i$ and $\boldsymbol{\delta}_i$ are normally distributed, and integrating over the LV, the joint marginal distribution of the observed outcome and exposures, $f(\boldsymbol{Y}_i^* \,|\, G_i = g, \boldsymbol{Z}_i, \boldsymbol{W}_i; \boldsymbol{\theta})$, is a multivariate normal density, with moments given by $E(\boldsymbol{Y}_i^* \,|\, G_i = g, \boldsymbol{Z}_i, \boldsymbol{W}_i; \boldsymbol{\theta}) = \boldsymbol{\nu}_g^* + \boldsymbol{\Lambda}_g^*(\boldsymbol{\alpha}_{0,g} + \boldsymbol{W}_i\boldsymbol{\alpha}_W) + \boldsymbol{\beta}_{Z,g}^* \boldsymbol{Z}_i$ and Var $(\boldsymbol{Y}_i^* \,|\, G_i = g, \boldsymbol{Z}_i, \boldsymbol{W}_i; \boldsymbol{\theta}) = \boldsymbol{\Lambda}_g^* \boldsymbol{\Phi}_g (\boldsymbol{\Lambda}_g^*)^\top + \boldsymbol{\Theta}_g^*$, where

$$\boldsymbol{\nu}_g^* = \begin{pmatrix} \beta_{0,g} \\ \boldsymbol{\nu}_g \end{pmatrix}, \quad \boldsymbol{\Lambda}_g^* = \begin{pmatrix} \boldsymbol{\beta}_{U,g} \\ \boldsymbol{\Lambda}_g \end{pmatrix},$$

$$\boldsymbol{\beta}_{Z,g}^* = \begin{pmatrix} \boldsymbol{\beta}_{Z,g} \\ 0_{p\times q} \end{pmatrix} \quad \text{and} \quad \boldsymbol{\Theta}_g^* = \begin{pmatrix} \sigma^2 & 0_{1\times p} \\ 0_{p\times 1} & \boldsymbol{\Theta}_g \end{pmatrix}.$$

The log likelihood of $\boldsymbol{\theta}$ is then, $\ell(\boldsymbol{\theta}) = \sum_{i=1}^N \sum_{g=0}^G \log f(\boldsymbol{Y}_i^* \,|\, G_i = g, \boldsymbol{Z}_i, \boldsymbol{W}_i; \boldsymbol{\theta})$. Parameter estimates are obtained by maximizing $\ell(\boldsymbol{\theta})$, or equivalently, solving the score equations $S(\boldsymbol{\theta}) = \sum_{i=1}^N \sum_{g=0}^G S_i = 0$, where $S_i = \partial \log f(Y_i^* \,|\, G_i = g, \boldsymbol{Z}_i, \boldsymbol{W}_i; \boldsymbol{\theta}) / \partial \boldsymbol{\theta}$ is the contribution of the ith observation to the score. Variances for parameters can be obtained by inverting the information matrix, $\boldsymbol{I}_\theta = -E(\partial^2 \ell(\boldsymbol{\theta}) / \partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top)$, or by computing robust variances $\widehat{var}_R(\widehat{\boldsymbol{\theta}}) = \boldsymbol{B}^{-1}\boldsymbol{A}\boldsymbol{B}^{-\top}$, where $\boldsymbol{A} = 1/N \sum_{i=1}^N \sum_{g=0}^G S_i S_i^\top$ and $\boldsymbol{B} = 1/N \sum_{i=1}^N \sum_{g=0}^G \partial S_i / \partial \boldsymbol{\theta}$.

3.2 *Shrinkage Estimation*

Shrinkage estimators have been used in outcome-dependent sampling based studies as a way to balance bias and efficiency gains from assuming *G-E* independence while using a retrospective likelihood formulation of the model. Shrinkage approaches will enhance efficiency only when the retrospective model improves efficiency. A typical formulation will factorize the retrospective likelihood $p(G, E, Z \,|\, Y; \theta_1, \theta_2, \theta_3)$ as $p(Y \,|\, G, E, Z; \theta_1) p(G \,|\, E, Z; \theta_2) p(E, Z; \theta_3) / p(Y; \theta_1, \theta_2, \theta_3)$ with θ_1 being the outcome model parameters, θ_2 and θ_3 describing the *G-E* dependence and exposure–covariate associations, respectively, and the *G-E* association reflected through the term $p(G \,|\, E, Z; \theta_2)$. Because $p(Y; \theta_1, \theta_2, \theta_3) = \sum_{G,E,Z} p(Y \,|\, G, E, Z; \theta_1) p(G \,|\, E, Z; \theta_2) p(E, Z; \theta_3)$ (in the denominator) depends on the specification of $p(G \,|\, E, Z; \theta_2)$, the MLE of θ_1 depends on the assumed model for the *G-E* association conditional on covariates. It has been shown (Chatterjee and Carroll, 2005) that assuming conditional *G-E* independence, $p(G \,|\, E, Z; \theta_2) = p(G \,|\, Z; \theta_2)$, in case–control studies leads to large efficiency gain for estimating the $G \times E$ interaction parameter in the outcome model $p(Y \,|\, G, E, Z; \theta_1)$. However, under violation of the independence assumption, these estimators are biased. Shrinkage estimators then arise as a weighted average of two estimators: one obtained under dependence and the other obtained under independence; the weights are chosen in a data-adaptive fashion and reflect the uncertainty around the conditional *G-E* association (Mukherjee and Chatterjee, 2008). Chen et al. (2009) propose a class of shrinkage estimators applicable, in principle, to estimation problems beyond $G \times E$ effects. Shrinkage estimates can be motivated from an Empirical Bayes (EB) perspective.

In a cohort or cross-sectional study, maximization of the joint likelihood with respect to outcome model parameters, even from modeling the joint distribution $p(Y, G, E \,|\, Z; \theta_1, \theta_2, \theta_3)$, will be independent of the specification of $p(G, E \,|\, Z; \theta_2, \theta_3)$. Because of the lack of outcome dependent sampling, and thus absence of conditioning on Y, the log likelihood becomes a sum of two terms that can be maximized separately.

However, the proposed LV framework allows us to incorporate constraints on the *G-E* association (Section 2.2), raising the possibility of gaining efficiency. In the LV framework, estimation of *G-E* model parameters cannot be completely disentangled from estimation of outcome model parameters due to integration of the likelihood over the LV. The extent of efficiency and power gains for testing of outcome parameters due to such constraints depends on design and effect size settings that we investigate in our simulation study. Although extensive work exists on using *G-E* independence assumptions in case–control studies, this article is the first to propose and study the implications such assumptions in an LV setting.

We follow Chen et al. (2009) to describe how estimates obtained under assumptions $A0$ and $A1 - A3$ can be combined. Denoting the two estimates by $\widehat{\boldsymbol{\theta}}_{A0}$ and $\widehat{\boldsymbol{\theta}}_{A*}$,

$$\widehat{\boldsymbol{\theta}}_{\text{shrink}} = \widehat{\boldsymbol{\theta}}_{A*} + K_{\text{MV}}(\widehat{\boldsymbol{\theta}}_{A0} - \widehat{\boldsymbol{\theta}}_{A*}) \tag{5}$$

is the shrinkage estimator, where A^* is any of $A1 - A3$ and shrinkage weights are $K_{\text{MV}} = \widehat{V}(\widehat{V} + \widehat{\boldsymbol{\psi}}\widehat{\boldsymbol{\psi}}^\top)^{-1}$, with $\widehat{\boldsymbol{\psi}} = \widehat{\boldsymbol{\theta}}_{A0} - \widehat{\boldsymbol{\theta}}_{A*}$ and $\widehat{V}$ is the estimated asymptotic covariance matrix of $\widehat{\boldsymbol{\psi}}$. Alternatively, one may use a diagonal weight matrix K_{CW}, where the kth diagonal element is $\widehat{V}_k/(\widehat{V}_k + \widehat{\psi}_k^2)$; $\widehat{V}_k$ being the kth diagonal element of $\widehat{V}$ and $\widehat{\psi}_k$ the kth component of $\widehat{\boldsymbol{\psi}}$. Choosing K_{CW} leads to "component-wise (CW)" shrinkage, because the weights used for a given component of $\widehat{\boldsymbol{\theta}}_{\text{shrink}}$ depend only on the variance and bias related to that component; we call these estimates EB_{CW}. In contrast, using K_{MV} leads to so-called multivariate (MV) shrinkage (Chen et al., 2009), which we refer to as EB_{MV}. In both cases we use superscripts to denote which estimates were combined, e.g., EB^{03}_{MV} denotes combination of estimates obtained under assumptions $A0$ and $A3$. Note that (5) is only defined for parameters that are common to both models, and we have slightly abused notation by using $\widehat{\boldsymbol{\theta}}_{A*}$ in (5) to only represent the subset of parameters that are equivalent to those in $\widehat{\boldsymbol{\theta}}_{A0}$.

Additional considerations about shrinkage estimators are worth mentioning. First, CW shrinkage may be desirable in terms of efficiency gain, compared to multivariate shrinkage in small samples, because large sampling error in the off-diagonals of $\widehat{V}$ undermines the potential efficiency gain from multivariate shrinkage (Chen et al., 2009). Second, the form of the weights imply that EB_{MV} estimates will be more prone to favoring the more flexible models. To see this, note that the component-wise shrinkage weights (the kth diagonal element of K_{CW}) can be rewritten as $1/(1 + \chi_k^2)$, where $\chi_k^2 = \widehat{\psi}_k^2/\widehat{V}_k$ is the ratio of the squared difference in the kth parameter between the two models divided by variance of the difference. The MV shrinkage weights can be similarly rewritten: $1/(1 + \chi^2)$ where $\chi^2 = \widehat{\boldsymbol{\psi}}^\top \widehat{V}^{-1}\widehat{\boldsymbol{\psi}}$. χ_k^2 and χ^2 can be interpreted as a bias–variance ratios; when they are smaller than one, the *EB* estimates will lean toward the simpler model. In contrast to CW shrinkage, the MV shrinkage weight is the same for all parameters, but the ratio χ^2 is a weighted sum of all the bias–variance ratios for all model parameters. Hence, in MV shrinkage, a given parameter might be shrunk given not only its own bias–variance ratio χ_k^2 but also given bias–variance ratios for other parameters. Furthermore, χ_k^2 has expectation (approximately) 1 when independence holds, whereas χ^2 has expectation equal to the number of model parameters estimated with assumption $A0$. Hence, MV shrinkage weights $1/(1 + \chi^2)$ will almost always be small, leading to EB estimates closer to those from more flexible models. Third, the weights summarize information about model fit in the sense of comparing differences in estimated parameters. If the models fit equally well, then corresponding parameters would likely be similar. Large differences in corresponding parameters indicate a poorer fitting model (e.g., constrained model). Hence, shrinkage estimates can help assess model adequacy, with smaller weights for the constrained model implying the more flexible model is preferred. Finally, both CW and MV shrinkage estimators will asymptotically converge to those from the more flexible model.

Chen et al. (2009) provide formal arguments to derive the variance for $\widehat{\boldsymbol{\theta}}_{\text{shrink}}$. Heuristically, the variance can be obtained by treating $\widehat{\boldsymbol{\theta}}_{\text{shrink}}$ as a function of two random variables, $\widehat{\boldsymbol{\theta}}_{A0}$ and $\widehat{\boldsymbol{\theta}}_{A*}$, with joint covariance matrix $\boldsymbol{\Sigma} = \text{Var}((\widehat{\boldsymbol{\theta}}_{A0}^\top, \widehat{\boldsymbol{\theta}}_{A*}^\top)^\top)$. Letting $h(\boldsymbol{\theta}_{A0}, \boldsymbol{\theta}_{A*}) = \boldsymbol{\theta}_{A*} + V(V + \boldsymbol{\psi}\boldsymbol{\psi}^\top)^{-1}\boldsymbol{\psi}$, with $\boldsymbol{\psi} = \boldsymbol{\theta}_{A0} - \boldsymbol{\theta}_{A*}$, and employing the multivariate Delta theorem, then $\text{Var}(\widehat{\boldsymbol{\theta}}_{\text{shrink}}) \approx \boldsymbol{H}^\top \widehat{\boldsymbol{\Sigma}} \boldsymbol{H}$, where $\boldsymbol{H} = \partial h(\boldsymbol{\theta}_{A0}, \boldsymbol{\theta}_{A*})/\partial(\boldsymbol{\theta}_{A0}^\top, \boldsymbol{\theta}_{A*}^\top)\,|_{\boldsymbol{\theta}_{A0}=\widehat{\boldsymbol{\theta}}_{A0}, \boldsymbol{\theta}_{A*}=\widehat{\boldsymbol{\theta}}_{A*}}$. Matrix $\widehat{\boldsymbol{\Sigma}} = \boldsymbol{D}^{-1}\boldsymbol{C}\boldsymbol{D}^{-\top}$ is constructed using the sandwich-variance formula where $\boldsymbol{C} = 1/N \sum_{i=1}^N P_i P_i^\top$ and $\boldsymbol{D} = 1/N \sum_{i=1}^N \partial P_i/\partial \boldsymbol{\theta}$, and $P_i = (S_{A0,i}^\top S^\top{}_{A*,i})^\top$ is a stacked vector of likelihood score contributions from each model.

3.3 *Simpler Estimation Strategies*

Instead of positing a latent exposure model, (1)–(4), one may fit separate multiple linear regression models (MLR) on each exposure measure, or one regression on their first principal component (PCA), and its interactions with *G*. We include these simple approaches in Sections 4 and 5.

4. Simulation Studies

We conducted a small-scale simulation study to examine the finite sample properties of estimators under various settings of the true data generating model using $l = 1$, $p = 4$, and two genetic classes. Genetic class was generated as a binary variable with prevalence 0.2, similar to our data example. Because there are only two gene classes and one latent exposure, in this section (and Section 5) we denote the gene effect among unexposed, the exposure effect among wild types, and the interaction parameters as β_G, β_U, $\beta_{G\times U}$. We investigate the estimators' properties under two scenarios of the *G-E* association: independence ($A0$) and dependence ($A3$). We used either $\beta_U = \beta_G = \beta_{G\times U} = 0$ or $\beta_U = 1$, $\beta_G = \beta_{G\times U} = 2$ (i.e., standardized effects of 0.2, 0.4, 0.4, respectively, because outcome variance was $\sigma^2 = 5^2$). See Supplementary Materials for full design.

Type I error: When *G* and *E* are independent, all approaches retain rejection probabilities (P(R)) of approximately 0.05 for tests at the 0.05 significance level (Table 1, scenario $A0$). When the data are generated under *G-E* dependence (Table 1, scenario $A3$), MLE estimates derived assuming independence ($A0$) have inflated type I error probabilities for $\widehat{\beta}_U$ and $\widehat{\beta}_G$. While EB_{CW} retain inflated type I error rates, EB_{MV} estimates do not.

Table 1

Bias, variance ratios (Var.R), MSE, and rejection probabilities (P(R)) for outcome model parameter estimates under two scenarios of true exposure model parameters. Outcome model parameters set at $\beta_U = \beta_G = \beta_{G\times U} = 0$, $\sigma^2 = 5$; sample size was $N = 350$, with 500[§] replicates.

Data	Est.	$\widehat{\beta}_U$				$\widehat{\beta}_G$				$\widehat{\beta}_{G\times U}$			
model	method[a]	Bias	Var.R[b]	MSE	P(R)	Bias	Var.R[b]	MSE	P(R)	Bias	Var.R[b]	MSE	P(R)
A0	*A*0	0.00	1(Ref)	0.15	0.063	−0.05	1(Ref)	0.46	0.048	−0.04	1(Ref)	0.65	0.036
	*A*1	0.00	(1.02)	0.16	0.059	−0.05	(1.05)	0.46	0.044	−0.04	(1.10)	0.69	0.038
	*A*2	0.00	(1.01)	0.15	0.059	−0.04	(1.11)	0.48	0.042	−0.05	(1.15)	0.72	0.036
	*A*3	0.00	(1.02)	0.15	0.061	−0.04	(1.12)	0.49	0.042	−0.04	(1.15)	0.74	0.044
	EB^{01}_{CW}	0.00	(1.02)	0.15	0.061	−0.05	(0.99)	0.46	0.046	−0.04	(0.99)	0.66	0.050
	EB^{02}_{CW}	0.00	(1.02)	0.15	0.069	−0.04	(1.02)	0.46	0.046	−0.04	(1.01)	0.66	0.050
	EB^{03}_{CW}	0.00	(1.02)	0.15	0.069	−0.05	(1.03)	0.46	0.044	−0.04	(1.04)	0.67	0.052
	EB^{01}_{MV}	0.00	(1.03)	0.16	0.063	−0.05	(1.02)	0.46	0.044	−0.04	(1.04)	0.69	0.046
	EB^{02}_{MV}	0.00	(1.03)	0.15	0.063	−0.04	(1.09)	0.48	0.044	−0.05	(1.08)	0.71	0.044
	EB^{03}_{MV}	0.00	(1.03)	0.15	0.063	−0.04	(1.09)	0.48	0.042	−0.04	(1.07)	0.73	0.055
	E1[c]	0.00	(0.23)	0.03	0.044	−0.04	(1.38)	0.64	0.052	0.00	(0.24)	0.15	0.048
	PCA[d]	0.00	(0.23)	0.03	0.063	−0.04	(1.02)	0.46	0.048	0.01	(0.24)	0.15	0.042
A3	*A*0	0.03	1(Ref)	0.18	0.126	−0.07	1(Ref)	0.77	0.142	0.03	1(Ref)	0.52	0.043
	A1	0.04	(2.59)	0.32	0.038	−0.10	(2.46)	1.15	0.063	0.02	(1.21)	0.66	0.050
	*A*2	0.03	(1.42)	0.18	0.036	−0.08	(2.07)	0.95	0.056	0.03	(1.08)	0.57	0.043
	*A*3	0.03	(1.27)	0.16	0.038	−0.09	(2.26)	1.04	0.047	0.04	(1.19)	0.63	0.050
	EB^{01}_{CW}	0.03	(1.38)	0.23	0.020	−0.07	(1.18)	0.85	0.101	0.02	(1.81)	0.56	0.007
	EB^{02}_{CW}	0.03	(0.83)	0.18	0.171	−0.07	(1.10)	0.80	0.119	0.03	(1.65)	0.52	0.011
	EB^{03}_{CW}	0.03	(0.84)	0.17	0.153	−0.08	(1.16)	0.82	0.106	0.03	(1.61)	0.53	0.018
	EB^{01}_{MV}	0.04	(2.59)	0.32	0.036	−0.10	(2.37)	1.15	0.074	0.02	(1.16)	0.66	0.050
	EB^{02}_{MV}	0.03	(1.39)	0.18	0.050	−0.08	(1.98)	0.95	0.056	0.03	(1.04)	0.57	0.054
	EB^{03}_{MV}	0.03	(1.25)	0.16	0.054	−0.09	(2.14)	1.03	0.059	0.04	(1.13)	0.63	0.061
	*E*1[c]	0.01	(0.28)	0.04	0.070	−0.00	(1.97)	0.87	0.052	−0.01	(0.23)	0.12	0.047
	PCA[d]	−0.01	(0.36)	0.04	0.038	−0.05	(1.57)	0.75	0.056	−0.01	(0.24)	0.12	0.061

[a] *A*0, *A*1, *A*2, *A*3 denote MLE; $A1^*$ not included because combining *A*0 and $A1^*$ resulted in singular. variance matrix $\boldsymbol{\Sigma}$ (see Section 5).
[b]Ratios of empirical variances, comparing to variance of *A*0.
[c]Multiple regression using one exposure marker, E_1.
[d]Multiple regression using first PCA of (E_1, E_2, E_3, E_4) as exposure marker.
[§] 4.6% (*A*0) and 7.0% (*A*3) data sets excluded to lack of convergence or unstable results, see Supplementary Materials for details.

Efficiency/power: When *G-E* independence holds (Table 2, scenario *A*0), gains in efficiency for $\widehat{\beta}_G$ and $\widehat{\beta}_{G\times U}$ estimated from *A*0 compared to *A*1–*A*3 are very clear: the variance ratio (Var.R) for $\widehat{\beta}_{G\times U}$ estimated under *A*3 versus *A*0 is 1.90. Efficiency gains translate to large power gains: power for $\widehat{\beta}_{G\times U}$ is 0.44 under *A*3 and 0.66 under *A*0. Compared to *A*3, EB^{03}_{CW} and EB^{03}_{MV} have lower variance ratios (1.46 and 1.74, respectively) and higher power (0.53 and 0.49, respectively). The PCA approach has power comparable to that from *A*0, despite the bias in $\widehat{\beta}_U$ and $\widehat{\beta}_{G\times U}$.

Bias: When *G-E* independence does not hold, all parameter estimates have large biases, except those from *A*3 and EB^{03}_{MV} (Table 2, scenario *A*3). Of the MLEs, bias is larger when *A*0 or *A*1 are assumed; further, *A*1 versus *A*0 does not result in uniformly less (absolute) bias for both β_U and $\beta_{G\times U}$. Simply relaxing the assumption of different mean and variance for the LV, but not the measurement model, may not be sufficient to reduce bias, and could in fact increase it.

EB_{CW} estimates are approximately half way between *A*0 and the more flexible models, although the exact distance varies depending on the magnitude of the coefficient. As such, they retain some of the bias of *A*0 estimates when *G-E* dependence exists. For EB^{01}_{CW}, the bias persists and is larger for β_U and $\beta_{G\times U}$ than the bias in *A*0 for these parameters. In contrast, EB^{02}_{MV} and EB^{03}_{MV} are generally closer to the more flexible model (see Section 3.2), and mostly eliminate the bias in *A*0.

As would be expected from the measurement error literature, parameter estimates using only *E*1 or PCA as the predictor in multiple regression analysis are biased, i.e., $\widehat{\beta}_U$ and $\widehat{\beta}_{G\times U}$ are attenuated. However, note that $\widehat{\beta}_G$ have large bias as well, due to the *G-E* dependence and the measurement error in *E*1 or the first PC in measuring exposure. Measurement error in one predictor (e.g., E_1) can induce bias in regression coefficients of covariates measured without error (e.g., *G*) that are correlated with the error-prone predictor (Budtz-Jørgensen et al., 2003a). The bias in $\widehat{\beta}_G$ can be in either positive or negative under the alternative hypothesis (Huang, Wang, and Cox, 2005), although estimates will be unbiased when there is no exposure effect ($\beta_U = \beta_{G\times U} = 0$, Table 1).

MSE: EB^{03}_{MV} estimates eliminate the bias in *A*0, but are less efficient; EB^{03}_{CW} retain some bias, but achieve smaller

Table 2

Percent bias, variance ratios (Var.R), MSE, and rejection probabilities (P(R)) for outcome model parameter estimates under two scenarios of true exposure model parameters. Outcome model parameters set at $\beta_U = 1$, $\beta_G = \beta_{G\times U} = 2$, $\sigma^2 = 5$; sample size was $N = 350$, with 500[§] replicates.

Data	Est.	$\widehat{\beta}_U$				$\widehat{\beta}_G$				$\widehat{\beta}_{G\times U}$			
model	method[a]	Bias%	Var.R[b]	MSE	P(R)	Bias%	Var.R[b]	MSE	P(R)	Bias%	Var.R[b]	MSE	P(R)
A0	*A*0	−1.3%	1(Ref)	0.17	0.74	3.4%	1(Ref)	0.47	0.85	−1.4%	1(Ref)	0.74	0.66
	*A*1	−1.0%	(1.02)	0.18	0.74	3.0%	(1.14)	0.54	0.78	1.3%	(1.18)	0.83	0.62
	*A*2	−1.0%	(1.03)	0.18	0.74	3.0%	(1.73)	0.74	0.63	3.1%	(1.74)	1.06	0.47
	*A*3	−0.7%	(1.04)	0.18	0.74	2.9%	(1.76)	0.75	0.63	3.1%	(1.90)	1.18	0.44
	EB^{01}_{CW}	−1.2%	(1.01)	0.17	0.73	3.2%	(1.08)	0.50	0.82	−0.3%	(1.08)	0.76	0.64
	EB^{02}_{CW}	−1.2%	(1.02)	0.17	0.74	3.4%	(1.45)	0.56	0.71	−0.5%	(1.36)	0.82	0.56
	EB^{03}_{CW}	−1.0%	(1.02)	0.17	0.75	3.3%	(1.47)	0.56	0.72	−1.1%	(1.46)	0.85	0.53
	EB^{01}_{MV}	−1.0%	(1.02)	0.18	0.73	3.0%	(1.12)	0.54	0.79	1.3%	(1.13)	0.83	0.64
	EB^{02}_{MV}	−1.0%	(1.03)	0.18	0.73	3.0%	(1.66)	0.73	0.64	3.1%	(1.64)	1.04	0.50
	EB^{03}_{MV}	−0.7%	(1.04)	0.18	0.73	2.9%	(1.67)	0.72	0.65	2.8%	(1.74)	1.14	0.49
	E1[c]	−65.4%	(0.22)	0.46	0.46	3.2%	(1.33)	0.49	0.83	−63.6%	(0.23)	1.79	0.41
	PCA[d]	−53.0%	(0.22)	0.31	0.74	3.3%	(0.94)	0.47	0.86	−51.5%	(0.23)	1.24	0.65
A3	*A*0	12.1%	1(Ref)	0.20	0.86	35.7%	1(Ref)	1.24	0.94	−27.6%	1(Ref)	0.92	0.48
	*A*1	45.6%	(2.57)	0.55	0.78	−18.7%	(2.57)	1.37	0.35	−42.0%	(1.24)	1.50	0.27
	*A*2	8.9%	(1.43)	0.19	0.79	10.3%	(3.79)	1.96	0.51	−13.5%	(1.86)	1.21	0.40
	*A*3	3.7%	(1.26)	0.17	0.80	−1.4%	(5.18)	2.14	0.41	0.9%	(2.65)	1.34	0.39
	EB^{01}_{CW}	32.8%	(2.51)	0.43	0.72	0.0%	(2.83)	1.23	0.44	−37.5%	(1.94)	1.30	0.19
	EB^{02}_{CW}	10.4%	(0.94)	0.19	0.88	28.4%	(3.47)	1.58	0.67	−21.3%	(1.98)	0.94	0.31
	EB^{03}_{CW}	7.4%	(0.95)	0.18	0.86	21.8%	(4.45)	1.56	0.57	−14.6%	(2.48)	0.93	0.30
	EB^{01}_{MV}	45.5%	(2.54)	0.55	0.77	−18.6%	(2.55)	1.36	0.37	−42.0%	(1.22)	1.51	0.27
	EB^{02}_{MV}	9.0%	(1.38)	0.19	0.79	10.7%	(4.03)	1.93	0.50	−13.7%	(1.97)	1.20	0.39
	EB^{03}_{MV}	3.8%	(1.23)	0.17	0.79	−0.7%	(4.89)	2.09	0.44	0.4%	(2.50)	1.31	0.41
	*E*1[c]	−63.5%	(0.27)	0.44	0.51	110.7%	(1.97)	5.59	0.99	−78.0%	(0.23)	2.59	0.23
	PCA[d]	−44.2%	(0.33)	0.23	0.78	37.3%	(1.52)	1.29	0.88	−63.7%	(0.23)	1.77	0.48

[a] *A*0, *A*1, *A*2, *A*3 denote MLE; $A1^*$ not included because combining *A*0 and $A1^*$ resulted in singular variance matrix $\boldsymbol{\Sigma}$ (see Section 5).
[b]Ratios of empirical variances, comparing to variance of *A*0.
[c]Multiple regression using one exposure marker, E_1.
[d]Multiple regression using first PCA of (E_1, E_2, E_3, E_4) as exposure marker.
[§] 3.6% (*A*0) and 2.4% (*A*3) data sets excluded to lack of convergence or unstable results, see Supplementary Materials for details.

mean squared error (MSE); hence, a better bias–efficiency tradeoff. For example, although EB^{03}_{CW} incurs 15% bias, its MSE is 0.93, in contrast to an MSE of 1.31 for EB^{03}_{MV}. EB_{CW} estimators achieve a better bias–variance compromise in small samples compared to EB_{MV}.

Additional simulation results for the case of null main effects and small interaction parameter: $\beta_U = 0$, $\beta_G = 0$, and $\beta_{G\times U} = 0.1$ demonstrate that efficiency gains in *A*0 versus *A*3 are still observed (Var.$R = 1.5$). Although to a smaller degree, bias in $\widehat{\beta}_{G\times E}$ persisted when incorrectly assuming *A*0 (11% versus 3% bias in *A*3).

Recommendation: For hypothesis testing alone, PCA approaches may be just as good as using a full LV model because they maintain type I error (Table 1) and have power comparable to *A*0 (Table 2). However, in terms of both bias and efficiency, using EB^{03}_{MV} is our recommended estimation strategy. Although EB^{03}_{MV} has slightly higher MSE than EB^{03}_{CW}, it yields unbiased estimates, provides better control of type I error, and has higher power than EB^{03}_{CW}.

5. Modeling Lead Exposure, Iron Metabolism Genes, and Birth Weight

We use data from the first ELEMENT cohort, where the following prenatal lead exposure biomarkers were collected on the mother and child: maternal blood lead levels at delivery and umbilical cord blood lead as well as maternal bone lead levels (patella and tibia) (Gonzalez-Cossio et al., 1997). Birth weight is the health outcome of interest in our analysis. To be included in this analysis, children had to be genotyped, and have measured birth weight and least one of the four prenatal exposure biomarkers ($N = 406$). Missing data on covariates was imputed five times (Raghunathan, Solenberger, and Van Hoewyk, 2002). Parameter estimates were obtained from the imputed data sets and combined across the imputed data sets according to standard formulae (Little and Rubin, 2002, p. 86).

Deleterious effects of prenatal lead exposure on birth weight have been demonstrated (Gonzalez-Cossio et al., 1997), and the main effects of lead exposure in this sample using the

Table 3

Outcome model parameter estimates, robust standard errors, and t-statistics obtained using MLE under assumptions A0–A3 and shrinkage-based estimates combining assumptions. Coefficients $\widehat{\beta}_U$ and $\widehat{\beta}_{G\times U}$ have been rescaled to represent changes in birth weight(g) associated with an increase of 10μgPb/g in patella bone mass. Models are adjusted for maternal age, parity, education, and marital status.

Est. method[a,b]	$\widehat{\beta}_U$	$se(\widehat{\beta}_U)$	T_U	$\widehat{\beta}_G$	$se(\widehat{\beta}_G)$	T_G	$\widehat{\beta}_{G\times U}$	$se(\widehat{\beta}_{G\times U})$	$T_{G\times U}$
$A0$	−80.59	29.97	$-2.69^{\dagger}$	−97.75	51.30	−1.91	92.42	63.50	1.46
$A1^{*}$	−79.51	29.79	$-2.67^{\dagger}$	−99.09	51.45	−1.93	91.92	62.56	1.47
$A1$	−98.44	39.54	$-2.49^{\dagger}$	−98.47	51.30	−1.92	103.09	52.71	$1.96^{\dagger}$
$A2$	−85.14	33.75	$-2.52^{\dagger}$	−98.83	51.42	−1.92	90.66	53.57	1.69
$A3$	−86.62	35.01	$-2.47^{\dagger}$	−100.74	51.41	$-1.96^{\dagger}$	105.00	49.24	$2.13^{\dagger}$
$\text{EB}_{\text{CW}}^{01^{*}}$	−80.05	29.76	$-2.69^{\dagger}$	−98.02	51.31	−1.91	91.90	63.25	1.45
$\text{EB}_{\text{CW}}^{01}$	−91.12	40.15	$-2.27^{\dagger}$	−97.89	51.31	−1.91	101.95	61.06	1.67
$\text{EB}_{\text{CW}}^{02}$	−81.73	32.13	$-2.54^{\dagger}$	−97.90	51.24	−1.91	92.58	60.07	1.54
$\text{EB}_{\text{CW}}^{03}$	−82.40	33.04	$-2.49^{\dagger}$	−98.93	51.37	−1.93	94.63	62.10	1.52
$\text{EB}_{\text{MV}}^{02}$	−85.02	33.64	$-2.53^{\dagger}$	−98.80	51.40	−1.92	90.70	53.76	1.69
$\text{EB}_{\text{MV}}^{03}$	−86.28	34.66	$-2.49^{\dagger}$	−100.58	51.39	$-1.96^{\dagger}$	104.30	49.81	$2.09^{\dagger}$
E_1^{c}	−41.58	15.06	$-2.76^{\dagger}$	−191.50	71.07	$-2.69^{\dagger}$	60.59	31.19	1.94
PCA[d]	−46.58	15.34	$-3.04^{\dagger}$	−187.92	69.27	$-2.71^{\dagger}$	58.52	30.04	1.95

[a,b]Estimates from models $A0$, $A1$, and $A1^{*}$, were not sufficiently distinguishable from each other; hence $\text{EB}_{\text{MV}}^{01}$ and $\text{EB}_{\text{MV}}^{01*}$ could not be obtained (Section 5).

[c]Multiple regression using patella lead, E_1.

[d]Multiple regression using first PCA of E_1, E_2, E_3, E_4 as exposure marker.

† p-value < 0.05.

LV model are significant ($\widehat{\beta}_U = -54.12, se(\widehat{\beta}_U) = 25.1, T_U = -2.16$, Supplementary Materials Table 2). However, individuals with at least one iron metabolism gene variant may be protected against reduced birth weight due to lead exposure (Cantonwine et al., 2010). However, because iron metabolism genes appear to up-regulate iron *and* lead absorption (Hopkins et al., 2008), there may be dependence between genotype status and lead exposure. We use two SNPs related to iron metabolism, variants of the hemochromatosis gene (C282Y and H63D), and dichotomize genotype into wild type for both (aa and bb) or variant for any (Aa, AA, bB, or BB); both SNPs were in Hardy–Weinberg equilibrium, and 83 participants (20%) were classified as variants.

Increasing lead exposure among wild types is associated with decreased birth weight (negative $\widehat{\beta}_U$ in Table 3). The largest point estimate is obtained under assumption $A1$, whereas the lowest is obtained using MLR with the observed patella lead measure as the exposure marker. Such large attenuation in the MLR estimate is due to measurement error of patella lead in capturing prenatal exposure. Similarly, the MLR effect estimated from a PCA-derived exposure summary is attenuated, consistent with the simulation studies. Among the MLE and *EB* estimates, those obtained from assumptions $A0$ and $A1^{*}$ have the smallest standard errors, that increase with increasing flexibility of the model as expected. The PCA and MLR standard errors are much smaller, and therefore, even though the point estimates are also largely attenuated, the test statistics are similar to those for the MLE estimates. Component-wise shrinkage estimates are approximately halfway between those from $A0$ and those from the more flexible models $A1^{*}$–$A3$. $\text{EB}_{\text{CW}}^{01}$ is closer to the estimates obtained assuming $A1$, but EB_{CW}^{02} and EB_{CW}^{03} are closer to the estimates from $A0$ than from those obtained with $A2$ or $A3$. In contrast, EB_{MV}^{02} and EB_{MV}^{03} are closer to the estimates from $A2$ or $A3$ compared to those from $A0$. CW shrinkage favors simpler models because it only trades off bias–variance in one parameter at a time, whereas MV shrinkage favors the more flexible model because it simultaneously considers differences in all model parameters (Section 3.2).

Estimates and standard errors for $\widehat{\beta}_G$ are fairly constant across *G-E* assumptions and *EB* estimates. However, $\widehat{\beta}_G$ from MLR and PCA are much higher (more negative) than those from MLE. This can be due to bias arising due to exposure–gene correlation and exposure measurement error (Budtz-Jørgensen et al., 2003a; Huang et al., 2005).

Being variant for iron metabolism genes is protective against reduced birth weight due to lead exposure ($\widehat{\beta}_{G\times U}$ are positive), as hypothesized (Cantonwine et al., 2010). Whereas assumptions about *G-E* independence did not ultimately alter the conclusions for the main effect among wild types (e.g., all t-test statistics $T_U < -2.4$ in all assumptions), conclusions about $\widehat{\beta}_{G\times U}$ are impacted by the assumed *G-E* model. The most flexible model yields largest estimated effects (105.0 g with $A3$ versus 92.4 g with $A0$) and the standard error is lower (49.2 in $A3$ versus 63.5 in $A0$), resulting in t-statistics for the $G \times U$ effect as low as $T_{G\times U} = 1.46$ ($A0$) and as large as 2.13 ($A3$). This is likely due to a higher degree of overall model residual variance explained in $A3$ due to an increased number of parameters (i.e., lowest −2 log likelihood, Table 4).

Differences in MLE estimates for the outcome parameters can be largely explained by a few key differences exposure model parameters by genotype (Table 4). While there is little difference in average exposure levels between wild types

Table 4
Exposure model parameter estimates and robust standard errors obtained under assumptions A0–A3 described by Figure 1

	A0	*A1**	*A1*	*A2*	*A3*
Model for LV	Est.(SE)	Est.(SE)	Est.(SE)	Est.(SE)	Est.(SE)
α_0	1.536 (0.078)	1.519 (0.086)	1.526 (0.081)	1.506 (0.087)	1.506 (0.086)
γ_g		0.084 (0.178)	0.050 (0.182)	0.148 (0.207)	0.148 (0.208)
$\Phi_{g=0}$	1.230 (0.312)	1.253 (0.317)	0.835 (0.218)	1.090 (0.306)	1.056 (0.333)
$\Phi_{g=1}$			1.645 (0.489)	1.440 (0.460)	2.112 (0.916)
Measurement model					
Estimates for wild types					
ν_2	1.031 (0.053)	1.021 (0.057)	1.024 (0.056)	1.040 (0.057)	1.040 (0.056)
ν_3	1.759 (0.024)	1.757 (0.025)	1.758 (0.025)	1.756 (0.027)	1.756 (0.027)
ν_4	2.043 (0.023)	2.041 (0.023)	2.042 (0.023)	2.034 (0.026)	2.034 (0.026)
λ_2	0.553 (0.126)	0.542 (0.124)	0.680 (0.136)	0.527 (0.136)	0.558 (0.157)
λ_3	0.140 (0.036)	0.138 (0.035)	0.156 (0.036)	0.129 (0.041)	0.131 (0.043)
λ_4	0.125 (0.032)	0.124 (0.032)	0.140 (0.033)	0.114 (0.037)	0.116 (0.039)
Θ_{11}	1.172 (0.283)	1.148 (0.289)	1.401 (0.219)	1.236 (0.262)	1.258 (0.311)
Θ_{22}	0.710 (0.099)	0.717 (0.097)	0.624 (0.103)	0.671 (0.096)	0.627 (0.104)
Θ_{33}	0.218 (0.017)	0.218 (0.017)	0.218 (0.017)	0.218 (0.017)	0.217 (0.019)
Θ_{44}	0.194 (0.015)	0.194 (0.015)	0.194 (0.015)	0.195 (0.015)	0.197 (0.017)
Θ_{34}	0.168 (0.014)	0.168 (0.014)	0.168 (0.014)	0.168 (0.014)	0.168 (0.016)
Estimates for variants[a]					
ν_2				0.880 (0.156)	0.920 (0.139)
ν_3				1.746 (0.059)	1.751 (0.056)
ν_4				2.053 (0.055)	2.056 (0.051)
λ_2				0.768 (0.162)	0.496 (0.205)
λ_3				0.180 (0.058)	0.142 (0.063)
λ_4				0.161 (0.053)	0.136 (0.058)
Θ_{11}					0.612 (0.809)
Θ_{22}					1.077 (0.259)
Θ_{33}					0.226 (0.040)
Θ_{44}					0.183 (0.032)
Θ_{34}					0.171 (0.033)
Model fit criteria[b],(criterion for good fit)					
Number of parameters[c]	23	24	25	31	36
−2LL (smaller is better)	12,302.8	12,302.6	12,296.6	12,291.4	12,287.0
AIC (smaller is better)	12,348.8	12,350.6	12,346.5	12,353.4	12,359.0
BIC (smaller is better)	12,440.9	12,446.7	12,446.6	12,477.6	12,503.2
CFI[d] (>.95)	0.963	0.962	0.970	0.968	0.967
TLI[d] (>.95)	0.961	0.959	0.967	0.962	0.958
RMSEA[d] (<.05)	0.041	0.042	0.038	0.041	0.043

[a] Parameters estimates for variants are the same for as for wild types unless shown here.
[b] For the exposure and outcome model combined, underlined values denote better fitting model.
[c] Including outcome model parameters.
[d] See Bentler and Hu (1995) for definitions.

and variants (small $\widehat{\gamma}_g$), the variance of the LV is twice as high among variants ($\widehat{\Phi}_{g=1} = 2.11$) than among wild types ($\widehat{\Phi}_{g=0} = 1.05$). Residual variances for E_1 and E_2, Θ_{11} and Θ_{22}, also appear to differ between genotypes (51% and 72% difference, respectively), as does λ_4 (17% difference). This deserves further study−e.g., differences in Θ_{11} and Θ_{22} might be due to maternal genotypes, which are inherently correlated to infant genotype. Such investigation is out of the scope of the current work, but this finding highlights the utility of LV models in elucidating potential biological pathways.

In this example, implementing multivariate shrinkage was possible only for combining estimates from *A*0 with those from *A*3 and *A*0 with *A*2 estimates. MV shrinkage using *A*0 and *A*1 (and *A*1*) estimates resulted in a numerically singular variance matrix $\mathbf{\Sigma} = \text{Var}((\widehat{\boldsymbol{\theta}}_{A0}^\top, \widehat{\boldsymbol{\theta}}_{A*}^\top)^\top)$, likely due to measurement model parameters being too similar (and correlated) when only making small changes in the LV model (3). Although in the example we implemented all approaches for exposition, and even though standard model fit criteria (Table 4) would point toward model *A*1 being a better model in this particular example, as a general strategy we prefer outcome model parameters estimated using the EB^{03}_{MV} approach. This approach avoids the potential for increased type I errors due to fitting multiple models before arriving at a final model, and minimizes bias in outcome model parameters that may persist due to differences in exposure model parameters associated to genotype that may not be declared "significantly different" due to lack of power.

6. Discussion

The presence of multiple correlated measures of exposure exacerbates existing challenges in $G \times E$ studies. The current article is the first step toward an integrated framework where LV models are used to reduce the dimensionality of the exposure measures, thereby limiting the number of tests made and boosting power. Due to the general model formulation, it is easy to accommodate measurement errors in predictors, a pervasive problem in environmental epidemiology, and reduce multicollinearity concerns. Furthermore, a major intuitive appeal of the LV approach is that it provides not only estimates of the disease model parameters, but also a clearer picture of the underlying *G-E* association and helps capture the essence of the scientific problem. For a genetic marker and exposure which may have a common metabolic pathway, this model is more meaningful to practitioners than a multiple regression model relating Y_i to G_i and $\boldsymbol{E}_i$, which is not informative about the association between G_i and $\boldsymbol{E}_i$.

Because of the flexibility afforded by LV models, one challenge is the potential for model misspecification. In this particular application of LV models, we described various specifications of the *G-E* association, and discussed how restrictions in the *G-E* model boost efficiency of the $G \times E$ associations, but may incur bias when such restrictions are incorrectly made. We proposed a strategy where one would fit a restricted model and the most flexible model afforded by the data, and then combine estimates based on shrinkage ideas. The proposed approach yields estimates that data adaptively compromise between bias and variance, and avoids having to fit and refit models until a best-fitting model is found. Alternatively, estimation could proceed in two stages. First, the most flexible model could be estimated, and genotype differences in exposure model parameters tested. In the second stage, parameters that were found to not differ by genotype would be constrained to be equal across genotypes. However, such two-stage approach would also suffer from inflated type I error (Mukherjee and Chatterjee, 2008). Yet another alternative, with a similar flavor to what we proposed here, is to average parameter estimates obtained under various *G-E* assumptions according to prior information of the *G-E* association (Li and Conti, 2009) or using model fit criteria as weights (Hjort and Claeskens, 2003). Further still, one could use LASSO or Ridge penalties to select which exposure model parameters vary by genotype (Leoutsakos et al., 2010). Lastly, extensions of the methods proposed could include using a continuous genetic risk score G, such that a larger number of genetic categories can be (indirectly) included without collapsing to a few categories due to limited sample size. Such extension may not be straightforward because the multiple group analysis used here would not apply. Compromise estimators like the ones presented have not been used in the LV modeling literature, but can be a tool to achieve improved modeling strategies and robustness in LV models in applications even beyond $G \times E$ studies.

It is possible that one may use the proposed approach for screening $G \times E$ effects in genome-wide interaction studies. In our simulation studies, the estimation procedure takes approximately 0.36 minutes per data set in a desktop computer with 3.2 GHz Intel processor and 1 GB RAM. In the advent of cluster and parallel computing the proposed approach is scalable to genome-wide studies. Nevertheless, if the intent is solely testing, and not estimation, the PCA approach may be suitable, because, as shown in the simulation studies, it had comparable power to the proposed shrinkage estimates, despite substantial bias. Employing dimension-reduction approaches to the environmental exposure data will reduce multiple testing problems because only one genome-wide scan would be needed, instead of one scan for each observed exposure. Our methods are particularly appealing to study $G \times E$ effects with a given environmental exposure and genetic subclasses defined through genes on a related metabolic pathway.

The availability of higher-dimensional genomic data, and multiple continuous or categorical outcomes point to several extensions of our work. General LV models encompass latent class models (Skrondal and Rabe-Hesketh, 2004); hence one could posit a latent class model for multiple genetic factors, $\boldsymbol{G}_i$, which borrows strength from multiple loci and can minimize the chance of false positives (Schumacher and Kraft, 2007). Recent proposals (Chatterjee et al., 2006) pose gene–gene interaction models based on an LV approach, and can be extended to reduce the dimension of gene–gene–environment interaction models. Similar to what we have done for the exposure model in the present article, a latent outcome model to summarize correlated multivariate or longitudinal outcome data $\boldsymbol{Y}_i$ can be proposed. One would summarize multivariate correlated outcomes by latent traits, i.e., express $\boldsymbol{Y}_i$ in terms of latent outcomes $\boldsymbol{f}_i$ (e.g., Budtz-Jørgensen et al., 2003b), and estimate model parameters for a regression of $\boldsymbol{f}_i$ on $\boldsymbol{U}_i$ and $\boldsymbol{G}_i$. When $\boldsymbol{Y}_i$ involves repeated measures over time (e.g., growth curves), the model for the observed multivariate vector $\boldsymbol{Y}_i$ for subject i, measured at multiple time points may contain a random slope and random intercept, which are inherently latent variables. The random effects can be modeled as dependent on $\boldsymbol{U}_i$ and $\boldsymbol{G}_i$ and other covariates, such that inferences on how exposure and genes modify growth rates can naturally be obtained. Moreover, multivariate observations reflecting LVs repeated over time (Roy and Lin, 2000), and time-to-event data (Proust-Lima et al., 2009) can be incorporated. In summary, extensions of the present model can involve summarization of all three data components: *Y*, *G*, *and E*.

7. Supplementary Materials

Supplementary Materials referenced in Section 4 are available under the Paper Information link at the `Biometrics` website `http://www.biometrics.tibs.org`.

Acknowledgements

The authors thank ELEMENT investigators for proving data for the example, as well as the following NIEHS grants that supported data collection: K23ES000381; P01 ES012874; P42 ES05947; R01 ES013744; R01 ES014930; R01 ES007821. The authors also acknowledge salary support from grants NSF DMS 1007494, NIEHS R01 ES016932 and R01 ES017022, and NIEHS/EPA 1-P20-SE018171-01.

References

Bentler, P. M. and Hu, L. T. (1995). Evaluating model fit. In *Structural Equation Modeling*, R. H. Hoyle (ed.), 76–99. London: Sage.

Bollen, K. A. (1989). *Structural Equations with Latent Variables*. New York: John Wiley & Sons.

Budtz-Jørgensen, E., Keiding, N., Grandjean, P., Weihe, P., and White, R. F. (2003a). Consequences of exposure measurement error for confounder identification in environmental epidemiology. *Statistics in Medicine* **22,** 3089–3100.

Budtz-Jørgensen, E., Keiding, N., Grandjean, P., Weihe, P., and White, R. F. (2003b). Statistical methods for the evaluation of health effects of prenatal mercury exposure. *Environmetrics* **14,** 105–120.

Cantonwine, D., Hu, H., Tellez-Rojo, M. M., Sánchez, B. N., Lamadrid-Figueroa, H., Ettinger, A. S., Mercado Garcia, A., Hernandez-Avila, M., and Wright, R. O. (2010). HFE gene variants modify the association between maternal lead burden and infant birthweight: A prospective birth cohort study in Mexico City, Mexico. *Environmental Health* **9,** 43. Available at http://www.ehjournal.net/content/9/1/43, accessed January 10, 2010.

Chatterjee, N. and Carroll, R. J. (2005). Semiparametric maximum likelihood estimation exploiting gene-environment independence in case-control studies. *Biometrika* **92,** 399–418.

Chatterjee, N., Kalaylioglu, Z., Moslehi, R., Peters, U., and Wacholder, S. (2006). Powerful multilocus tests of genetic association in the presence of gene-gene and gene-environment interactions. *American Journal of Human Genetics* **79,** 1002–1016.

Chen, Y. H., Chatterjee, N., and Carroll, R. J. (2009). Shrinkage estimators for robust and efficient inference in haplotype-based case-control studies. *Journal of the American Statistical Association* **104,** 220–233.

Dhungana, P., Eskridge, K. M., Baenziger, P. S., Campbell, B. T., Gill, K. S., and Dweikat, I. (2007). Analysis of genotype-by-environment interaction in wheat using a structural equation model and chromosome substitution lines. *Crop Science* **47,** 477–484.

Gonzalez-Cossio, T., Peterson, K. E., Sanin, L. H., Fishbein, E., Palazuelos, E., Aro, A., Hernandez-Avila, M., and Hu, H. (1997). Decrease in birth weight in relation to maternal bone-lead burden. *Pediatrics* **100,** 856–862.

Hjort, N. L. and Claeskens, G. (2003). Frequentist model average estimators. *Journal of the American Statistical Association* **98,** 879–899.

Hopkins, M. R., Ettinger, A. S., Hernandez-Avila, M., Schwartz, J., Tellez-Rojo, M. M., Lamadrid-Figueroa, H., Bellinger, D., Hu, H., and Wright, R. O. (2008). Variants in iron metabolism genes predict higher blood lead levels in young children. *Environmental Health Perspectives* **116,** 1261–1266.

Huang, G. H. and Bandeen-Roche, K. (2004). Building an identifiable latent class model with covariate effects on underlying and measured variables. *Psychometrika* **69,** 5–32.

Huang, L. S., Wang, H. K., and Cox, C. (2005). Assessing interaction effects in linear measurement error models. *Journal of the Royal Statistical Society Series C—Applied Statistics* **54,** 21–30.

Javaras, K. N., Hudson, J. I., and Laird, N. M. (2010). Fitting ACE structural equation models to case-control family data. *Genetic Epidemiology* **34,** 238–245.

Jeannie-Marie S. Leoutsakos, J. M. S., Bandeen-Roche, K., Garrett-Mayer, E., and Zandi, P. P. (2010). Incorporating scientific knowledge into phenotype development: Penalized latent class regression. *Statistics in Medicine* **30,** 784–798.

Khoury, M. J. and Wacholder, S. (2009). Invited commentary: From genome-wide association studies to gene-environment-wide interaction studies–challenges and opportunities. *American Journal of Epidemiology* **169,** 227–230; discussion 234–235.

Li, D. L. and Conti, D. V. (2009). Detecting gene-environment interactions using a combined case-only and case-control approach. *American Journal of Epidemiology* **169,** 497–504.

Little, R. J. A. and Rubin, D. B. (2002). *Statistical Analysis with Missing Data*, 2nd edition. Hoboken, New Jersey: John Wiley & Sons.

Mukherjee, B. and Chatterjee, N. (2008). Exploiting gene-environment independence for analysis of case-control studies: An empirical bayes-type shrinkage estimator to trade off between bias and efficiency. *Biometrics* **64,** 685–694.

Proust-Lima, C., Joly, P., Dartigues, J. F., and Jacqmin-Gadda, H. (2009). Joint modelling of multivariate longitudinal outcomes and a time-to-event: A nonlinear latent class approach. *Computational Statistics & Data Analysis* **53,** 1142–1154.

Qi, L., Ma, J., Qi, Q., Hartiala, J., Allayee, H., and Campos, H. (2011). Genetic risk score and risk of myocardial infarction in Hispanics. *Circulation* **123,** 374–380.

Raghunathan, T. E., Solenberger, P., and Van Hoewyk, J. (2002). *IVEware: Imputation and Variance Estimation Software User Guide*. Ann Arbor, Michigan: Survey Methodology Program, University of Michigan.

Rathouz, P., Van Hulle, C., Rodgers, J., Waldman, I., and Lahey, B. (2008). Specification, testing, and interpretation of gene-by-measured-environment interaction models in the presence of gene-environment correlation. *Behavioral Genetics* **38,** 301–315.

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

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.

Schumacher, F. R. and Kraft, P. (2007). A Bayesian latent class analysis for whole-genome association analyses: An illustration using the gaw15 simulated rheumatoid arthritis dense scan data. *BMC Proceedings* **1,** S112. Available at http://www.biomedcentral.com/1753-6561/1/S1/S112, accessed December 10, 2010.

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

Westland, J. C. (2010). Lower bounds on sample size in structural equation modeling. *Electronic Commerce Research and Applications* **9,** 476–487.

Received December 2010. *Revised May* 2011.
Accepted July 2011.