Published in final edited form as:

J Am Stat Assoc. 2008 March 1; 103(481): 201–211. doi:10.1198/016214507000000428.

Application of Multidimensional Selective Item Response Regression Model for Studying Multiple Gene Methylation in SV40 Oncogenic Pathways

Haiqun Lin [Associate Professor],

Division of Biostatistics, Yale University School of Medicine, New Haven, CT 06520
(haiqun.lin@yale.edu)

Ziding Feng,

Public Health Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, WA 98109

Yan Yu [Associate Professor],

Department of Quantitative Analysis and Operations Management, University of Cincinnati,
Cincinnati, OH 45221

Yingye Zheng [Assistant Member],

Public Health Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, WA 98109

Narayan Shivapurkar [Assistant Professor], and

Department of Pathology, University of Texas Southwestern Medical Center, Dallas, TX 75390

Adi F. Gazdar [Professor]

Department of Pathology, University of Texas Southwestern Medical Center, Dallas, TX 75390

Abstract

Alteration of gene methylation patterns has been reported to be involved in the early onsets of many human malignancies. Many exogenous risk factors, such as cigarette smoke, dietary additives, chemical exposures, radiation, and biologic agents including viral infection, are involved in the methylation pathways of cancers. We propose a multidimensional selective item response regression model to describe and test how a risk factor may alter molecular pathways involving aberrant methylation of multiple genes in oncogenesis. Our modeling framework is built on an item response model for multivariate dichotomous responses of high dimension, such as aberrant methylation of multiple tumor-suppressor genes, but we allow risk factors such as SV40 viral infection to alter the distribution of the latent factors that subsequently affect the outcome of cancer. We postulate empirical identification conditions under our model formulation. Moreover, we do not prespecify the links between the multiple dichotomous methylation responses and the latent factors, but rather conduct specification searches with a genetic algorithm to discover the links. Parameter estimation through maximum likelihood and specification searches in models with multidimensional latent factors for multivariate binary responses have become practical only recently, due to modern statistical computing development. We illustrate our proposal with the biological finding that simultaneous methylation of multiple tumor-suppressor genes is associated with the presence of SV40 viral sequences and with the cancer status of lymphoma/leukemia. We are able to test whether the data are consistent with the causal hypothesis that SV40 induces aberrant methylation of multiple genes in its oncogenic pathways. At the same time, we are able to evaluate the role of SV40 in the methylation pathway and to determine whether the methylation pathway is responsible for the development of leukemia/lymphoma.

Keywords

Biomarker; Causal pathway; Factor analysis; Genetic algorithm; Identification; Item response; Joint model; Latent variable; Specification search

1. INTRODUCTION

Alteration of gene methylation patterns constitutes an epigenetic activation–inactivation mechanism for growth regulatory genes responsible for cell proliferation control in the early events of many human malignancies (Jones and Laird 1999). Many exogenous risk factors, including cigarette smoke, dietary additives, occupational and environmental chemical exposures, and biologic agents (e.g., viral sequences), are involved in the methylation pathways of cancers. Aberrant methylation of CpG islands in promoter regions is a frequent method of silencing tumor-suppressor genes during the process of oncogenesis (Paulson, Fingerroth, Yates, and Speck 2002); for example, in malignant mesothelioma and gastric cancer, *RASSF1A* methylation is associated with virus infection of SV40 and Epstein–Barr Virus (EBV) (Toyooka et al. 2001; Dammann et al. 2003). A statistically significant association has been found between *MGMT* promoter hypermethylation and high levels of aflatoxin-DNA adduct in the development of hepatocellular carcinoma due to exposure of aflatoxin B-1 (Zhang et al. 2003). A significantly higher prevalence of P15 methylation has been found in histologically normal surgical margin epithelia of patients with head and neck squamous cell carcinomas with chronic smoking and drinking habits compared with nonsmokers and nondrinkers (Wong et al. 2003).

Simian virus 40 (SV40) is a potent oncogenic virus of rhesus monkey origin that seems to have spread to human beings through contamination of poliovirus stocks between 1955 and 1963, as well as by other means (Fisher, Weber and Carbone 1999). Besides mesotheliomas, brain tumors, and bone tumors, SV40 sequences also have been detected in hematologic cancers by several laboratories; their presence has been demonstrated in about 40% of lymphomas (Shivapurkar et al. 2002). SV40-positive mesotheliomas have been shown to have higher methylation frequencies than SV40-negative tumors, and SV40 sequence–containing tumors and those lacking such sequences demonstrate very different genetic changes, including methylation profiles (Toyooka et al. 2001). One objective of this research is to investigate the relationship among the presence of the viral sequences, gene methylation, and human lymphoma/leukemia.

In the data set analyzed in this article, aberrant methylations of the promoter regions of 13 known or suspected tumor-suppressor genes were examined in 90 cases of lymphoma/leukemia and 116 normal controls and cases of plasma cell dyscrasia (Shivapurkar et al. 2004). Pairwise correlation was detected among the status of aberrant methylation in each gene, the presence of SV40 sequence, and the cancer status (Table 1). Although we can simply regress cancer status on aberrant methylations (13 binary indicator variables) and the presence of SV40 sequence (also a binary indicator) through a logistic model, many coefficients have negative signs with high statistical significance due to the collinearity among the binary indicators. (The negative sign indicates the presence of the aberrant methylation in a gene is associated with decreased incidence of the cancer, which goes against our scientific evidence.) The multiplicity and overlaps in biological function of the genes under study further complicate the relationships. Our objectives were to evaluate (a) whether the association between SV40 and the cancer status is mediated through the multiple gene methylation, because the pairwise or multiway association does not warrant such a claim; (b) possible functional groups of the aberrantly methylated genes in the methylation pathway regardless of whether or not methylation mediates the path from SV40 to the cancer; (c) the contribution of each individual

gene in a functional group; (d) whether there exists any significant mechanism other than aberrant gene methylation in the pathway from SV40 to cancer. We are not able to achieve such goals simply by using the “usual” statistical approaches, such as chi-squared correlation study and logistic regression. The joint model that we developed will help reveal the possible mechanisms involved in the methylation pathway of cancer.

For the aforementioned purposes, we present a statistical model that we call *multidimensional selective item response regression* (MSIRR), which allows us to assess complex mediational mechanisms of the methylation pathway through the decomposition of the effects of SV40 viral sequences on the pathway. To discover the structure in the data under sound scientific restrictions, we apply a genetic algorithm (GA) of specification searches outlined by Marcoulides and Drezner (2001) for conventional linear factor models. These authors illustrated how a GA can help identify the data structure in a simple confirmatory factor analysis model based on five observed continuous variables. To the best of our knowledge, this article is the first work to implement a GA in a much more complicated model with latent variables for structured multivariate binary data of high dimension. Although the idea behind our starting model that we adopt in this article resembles that of multidimensional models for dichotomous item response data described by Reckase (1997), our model differs in that (a) it is fully identifiable; (b) both the mean vector and covariance matrix of latent factors in our model specification are allowed to depend on the input (i.e., whether a SV40 viral sequence is present); (c) the link between the latent factors and the dichotomous items in our MSIRR model is not prespecified before being fitted and tested with data, but is subject to searches by a GA under plausible scientific conditions; and (d) the trait associated with a latent factor may have a different interpretation if more items linked to the factor are added or removed, which is why we conduct a genetic search algorithm.

Some worthy aspects of our work from a statistical view point are that we provide a solution for model identification issues surrounding factor-/item response-type models that have not been systematically addressed in previous literature (see Sec. 3.2) and that we use a GA in a much more complex latent factor modeling framework. These aspects lead us to discover the biological data structure that would not be otherwise revealed by more conventional factor models (see Sec. 5.3). Our article also provides an illustration of how a factor-type model can be extended and used feasibly outside the fields of psychological measurements and educational testing. Another example of extending a conventional factor-type model was given by Dunson (2000), who developed a multivariate Bayesian latent trait model for mixed continuous and binary outcomes using a reproductive toxicity data; we adopt maximum likelihood estimation in this article.

The remainder of the article is organized as follows. In Section 2 we describe the methylation data set that provides a good resource for studying the gene methylation pathway in hematologic cancers. In Section 3 we specify our MSIRR model and present the identification conditions. In Section 4 we elaborate on estimation of the model and the specification searches through a GA. In Section 5 we describe the results from analysis of the methylation data. We conclude with a discussion in Section 6.

2. THE GENE METHYLATION STUDY

Blood cells from 90 cases of lymphoma/leukemia and 116 noncases were analyzed for SV40 viral sequences and promoter methylation of tumor-suppressor genes. SV40 tag sequences were found in 33 of 90 (36%) hematologic cancers and in 10 of 116 (8.6%) noncases. The promoter methylation status of 13 known or suspected tumor-suppressor genes was determined. The frequencies of methylation of the 13 genes in the hematologic malignancies varied from 11% for *RASSF1A* to 96% for *SHP1*. We found that presence of SV40 sequences

is significantly associated with aberrant methylation in all 13 genes listed except *P73*, with which it is marginally associated (Table 1). The frequency of detecting SV40 sequences is significantly higher in cases than in controls. Actually, SV40 sequences are significantly associated with leukemia, lymphoma, and both combined ($p = .015$, $<.001$, and $<.001$). Cancer cases have significantly higher frequencies of aberrant methylation in all genes except *P16*, *DcR1*, and *CRBP* (Table 1). Furthermore, association in aberrant methylation is detected for most pairs in the 13 genes.

But the detected pairwise association (although quite strong) between the presence of SV40 viral sequences, aberrant methylation in each of the genes, and leukemia/lymphoma does not validate the mediational effect of the aberrantly methylated genes. Neither does it provide information on relative importance of the genes in the SV40-cancer pathway. Due to the collinearity of the covariates, logistic regression of cancer status on the status of aberrant methylation in each of the genes and the presence of an SV40 sequence also does not allow us to untangle the multifaceted relationships. The model framework that we propose in next section will allow us to do this under plausible scientific assumptions.

3. STRUCTURE OF THE MULTIDIMENSIONAL SELECTIVE ITEM RESPONSE REGRESSION MODEL

Suppose that we have n independent samples labeled $i = 1, \dots, n$ and J genes of interest labeled $j = 1, \dots, J$. Let $Z_i = 1$ if a SV40 viral sequence is detected in sample i and 0 otherwise. Let $U_{ij} = 1$ if gene j in sample i is aberrantly methylated and 0 otherwise. Let $Y_i = 1$ if sample i is a case (lymphoma/leukemia) and 0 otherwise. Therefore, $Z_i, U_{i1}, \dots, U_{iJ}$, and Y_i consist of the observed data for sample i , we study the relationship among them through a joint model that we propose next.

3.1 Model Specification

Suppose that J genes can be divided into K functional groups of aberrant methylation that are represented by K latent factors and that SV40 sequences can affect the level of aberrant methylation in the K functional groups. Therefore, aberrant methylation in a gene is regulated through the functional groups. The aberrant methylation may mediate the path from SV40 to cancer, although other mechanisms may be involved as well. All of these features can be described in the joint model detailed later.

We use the following selective logit model to describe the extent to which gene j is aberrantly methylated in sample i :

$$\text{logit} \mu_{ij} = \beta_{0j} + \sum_{k=1}^K S_{jk} \beta_{jk} \eta_{ik}, \quad j=1, \dots, J, k=1, \dots, K, \quad (1)$$

where $\mu_{ij} = P\{U_{ij} = 1\}$, β_{0j} is the intercept of item location parameter; β_{jk} is the coefficient of factor loading for the k th latent factor η_{ik} , which is further modeled through a regression model (2) given next; and S_{jk} is the selection indicator taking value of 1 if aberrant methylation in gene j is affected by k th functional group and 0 otherwise. We impose the restriction that

$\sum_{k=1}^K S_{jk} \geq 1$ for all $j \in \{1, \dots, J\}$, because gene j will be excluded from any connection with other genes, SV40, and cancer status if $\sum_{k=1}^K S_{jk} = 0$. It should be noted that β_{jk} should only be estimated when $S_{jk} = 1$ and when the identification conditions postulated in Section 3.2.2 are satisfied. The relationship between K and J is specified in Section 3.2.2 to ensure identifiability.

$J_k(\leq J)$ genes can be involved in the k th functional group of aberrant methylated genes. We do not prespecify the gene items in the set $\{1, \dots, J_k\}$, but rather conduct specification searches to decide which gene items should belong to the set.

The aberrant methylation level in functional group k in sample i is modeled as

$$\eta_{ik} = \alpha_k Z_i + \varepsilon_{Zik}, \quad k=1, \dots, K, \quad (2)$$

where α_k is the mean level of aberrant methylation in functional group k among SV40 positive samples and the residual ε_{Zik} is assumed to be normally distributed with mean 0 and variance σ_{Zk}^2 , which depends on whether the sample is SV40 positive or negative. The covariance of η_{ik} and $\eta_{ik'}$ (or, equivalently, of ε_{Zik} and $\varepsilon_{Zik'}$) is assumed to be $\rho_{Zkk'}$, which also depends on Z . We require that the elements in the variance–covariance matrix of $(\eta_{i1}, \dots, \eta_{iK})$ are all freely estimated and impose identification conditions on other model parameters in next section.

The cancer status Y_i is modeled as

$$\text{logit}(EY_i) = \gamma_0 + \sum_{k=1}^K \gamma_k \eta_{ik} + \gamma_Z Z_i, \quad (3)$$

in which γ_0 is the intercept, the γ_k 's are coefficients associated with the k th latent factors, and γ_Z is the coefficient associated with Z_i , which represents the strength of the nonmethylation pathway in carcinogenesis induced by SV40 viral sequences.

In the foregoing specification, the aberrantly methylated genes may be related to different underlying functional groups that are described by the latent factors. The observed cancer outcome may be related directly to all functional groups in aberrant methylation. An overview of the structure of our model is given in Figure 1. $S_{jk} = 1$ indicates the involvement of gene j in functional group k . Given $S_{jk} = 1$, a test of $\beta_{jk} = 0$ indicates whether aberrant methylation of gene j is up-regulated or down-regulated by functional group k . A test of $\beta_k = 0$ indicates whether the aberrant methylation level of genes in functional group k is affected by SV40. The test of $\gamma_k = 0$ indicates whether the aberrant methylation in the k th functional group mediates the effect of SV40 on cancer outcome. The test of $\gamma_Z = 0$ indicates whether there exist any other nonmethylation mechanism(s) by which SV40 exerts its oncogenic effects.

The underlying variable approach (UVA) described by Moustaki, Jöreskog, and Mavridis (2004) is similar to our model in terms of item response with covariates and is supported by commercially available software, such as LISREL (Jöreskog and Sörbom 1999) and Mplus (Muthén and Muthén 2001); however, the links between items and factors in these models need to be prespecified, and the software requires further restriction on the variance–covariance matrix for the underlying latent responses. Both the underlying variables and the latent factors in our model can be interpreted as having certain traits. The term UVA is often associated with ordinal response and usually has probit link with thresholds; the item response model with latent factors is usually used in binary response.

3.2 Model Identification

Identification is essential for consistent estimation of model parameters and should be carefully studied for latent variable models due to the growing popularity of these complex models (Rabe-Hesketh and Skrondal 2001). There is a paucity of research regarding identification of

latent response models, as well as multidimensional item response models. Bollen (1989), Ihara and Kano (1995), and Rabe-Hesketh and Skrondal (2001) have illustrated some treatments with respect to identification of usual factor models and factor-structured logit/probit models.

In this section we study the sufficient and necessary identification conditions for the initial MSIRR model starting with all S_{jk} 's assumed to be 1. Identifying the MSIRR model requires that some of the S_{jk} 's be preset to 0 before conducting a model specification search and obtaining parameter estimates. The identification conditions thus derived will be sufficient for the models with some additional S_{jk} 's set to 0 by the specification search described in Section 4.2, because the starting model has the maximum number of unknown parameters β_{jk} .

3.2.1 Rotational Nonidentification—The second term in (1) for all of the J genes and the second term in (3) can be jointly written as $\mathbf{B}\boldsymbol{\eta}_i$, where $\mathbf{B}=(\beta_1^T, \dots, \beta_J^T, \beta_{J+1}^T)^T$, with $\{\beta_j = (\beta_{j1}, \dots, \beta_{jK})^T; j = 1, \dots, J\}$ and $\beta_{J+1} := \boldsymbol{\gamma} = (\gamma_1, \dots, \gamma_K)^T$. Because $\boldsymbol{\eta}_i = (\eta_{i1}, \dots, \eta_{iK})^T$ in (2) can be considered to have a multivariate normal distribution with mean $Z_i\boldsymbol{\alpha}$ and variance–covariance $\boldsymbol{\Sigma} + Z_i\boldsymbol{\Theta}$ with $\boldsymbol{\Sigma}$ being the variance–covariance matrix when $Z_i = 0$ and $\boldsymbol{\Theta}$ being the difference variance–covariance matrix of $\boldsymbol{\eta}_i$ between those $Z_i = 1$ and those $Z_i = 0$, $\mathbf{B}\boldsymbol{\eta}_i$ is multinormally distributed with mean $Z_i\mathbf{B}\boldsymbol{\alpha}$ and covariance $\mathbf{B}(\boldsymbol{\Sigma} + Z_i\boldsymbol{\Theta})\mathbf{B}^T$.

Without imposing additional identification conditions on the model, we can show that the $\{\mathbf{B}, \boldsymbol{\Sigma}, \boldsymbol{\alpha}, \boldsymbol{\Theta}\}$'s are globally unidentifiable because for a given solution $\{\mathbf{B}, \boldsymbol{\Sigma}, \boldsymbol{\alpha}, \boldsymbol{\Theta}\}$, any $\{\mathbf{B}^*, \boldsymbol{\Sigma}^*, \boldsymbol{\alpha}^*, \boldsymbol{\Theta}^*\}$ satisfying

$$\begin{aligned} \mathbf{B}^* &= \mathbf{B}\mathbf{T}, & \boldsymbol{\Sigma}^* &= \mathbf{T}^{-1}\boldsymbol{\Sigma}\mathbf{T}^{T,-1}, \\ \boldsymbol{\alpha}^* &= \mathbf{T}^{-1}\boldsymbol{\alpha}, & \boldsymbol{\Theta}^* &= \mathbf{T}^{-1}\boldsymbol{\Theta}\mathbf{T}^{T,-1} \end{aligned} \quad (4)$$

for a given $K \times K$ nonsingular square matrix $\mathbf{T}$ will also satisfy

$$\begin{aligned} \mathbf{B}\boldsymbol{\Sigma}\mathbf{B}^T &= \mathbf{B}^*\boldsymbol{\Sigma}^*\mathbf{B}^{*,T}, \\ \mathbf{B}\boldsymbol{\alpha} &= \mathbf{B}^*\boldsymbol{\alpha}^*, \\ \mathbf{B}\boldsymbol{\Theta}\mathbf{B}^T &= \mathbf{B}^*\boldsymbol{\Theta}^*\mathbf{B}^{*,T}. \end{aligned}$$

Thus different solutions can be produced by rotating $\mathbf{B}$, $\boldsymbol{\Sigma}$, $\boldsymbol{\alpha}$, and $\boldsymbol{\Theta}$ with $\mathbf{T}$. Even when the first factor loadings $\{\beta_{1k}, k = 1, \dots, K\}$ for each of the K factors are set to 1, the model with $K \geq 2$ is still not identified, because this imposes only K restrictions, whereas there are K^2 free elements in $\mathbf{T}$.

The foregoing identification issues are similar to the multidimensional two-parameter logistic (MLPL) model in which $Z_i = 0$ for all i 's. Due to the lack of systematic investigation of identifiability in multivariate factor models for categorical data and the limitation of statistical computational capacity, many published works assume an identity matrix for variance–covariance matrix of the latent factors.

3.2.2 Identifiability Conditions—Rabe-Hesketh and Skrondal (2001) used empirical underidentification to indicate the situation when the model is identified only from third- or higher-order moments; for example, the multivariate logit-normal model (Coull and Agresti 2000) having the dimension in multivariate random intercepts equal to the number of observations in a cluster is empirically underidentified, although still theoretically identified (Rabe-Hesketh and Skrondal 2001). A theoretically identified but empirically underidentified model often has near-singular information matrix at the maximum likelihood estimates,

because the third- and higher-order moments of the latent variables have a high sampling variability, and the observed dichotomized responses retain little information about the higher-order moments. Theoretical identification of the probit-normal version of the model ensures empirical identification for both probit-normal and logit-normal versions of the model.

Similarly, empirical identification of our MSIRR model will be ensured for either the logit or probit link in (1) and in (3) if the probit version of the model is theoretically identified. The information in our MSIRR model under probit link is contained in the first- and second-order moments of the underlying normalized latent responses. Thus, for a probit version of MSIRR model, we have the following identifying equations:

$$\begin{aligned} \mu_{ij} &= P(U_{ij}=1|Z_i) \\ &= \Phi \left[\frac{\beta_{0j} + \beta_{j1}\alpha_1 Z_i + \dots + \beta_{jK}\alpha_K Z_i}{\sqrt{\beta_j^T (\Sigma + Z_i \Theta) \beta_j + 1}} \right], \\ & \quad j=1, \dots, J, \end{aligned} \quad (5a)$$

$$\begin{aligned} \mu_{i,J+1} &:= E(Y_i) = P(Y_i=1|Z_i) \\ &= \Phi \left[\frac{\gamma_0 + \gamma_1 \alpha_1 Z_i + \dots + \gamma_K \alpha_K Z_i + \gamma_Z Z_i}{\sqrt{\beta_{J+1}^T (\Sigma + Z_i \Theta) \beta_{J+1} + 1}} \right], \end{aligned} \quad (5b)$$

and

$$\begin{aligned} \text{cov} \left(h(\mu_{ij^*}), h(\mu_{ij^{**}}) \right) &= \beta_{j^*}^T (\Sigma + Z_i \Theta) \beta_{j^{**}} + \mathbf{I}[j^* = j^{**}], \\ j^*, j^{**} &\in \{1, \dots, J+1\}, j^* \leq j^{**}, \end{aligned} \quad (5c)$$

where (5a) and (5b) are the marginal probabilities, β_{J+1} in (5b) is defined in Section 3.2.1 as γ , (5c) is the correlation matrix of the underlying latent responses under probit link, h in (5c) denotes probit link, and $\mathbf{I}$ in (5c) is an indicator variable.

Before presenting the following empirical identification conditions, we introduce the terms for item complexity with respect to the association of an item with the latent factors. An item j is said to be *factorially simple* if it is linked only to the factor k and not to any other factors, so $\beta_{jk} = 1$, but $\beta_{jk'} = 0$ for $k' = 1, \dots, k-1, k+1, \dots, K$, which implies $S_{jk'} = 0$. Correspondingly, an item with nonzero loadings on more than one factor is *factorially complex*. The postulated sufficient and necessary conditions for empirical identification in the sense of Rabe-Hesketh and Skrondal for our initial MSIRR model are (a) at least one factor loading for each factor k is set to 1, (b) at least one item is factorially simple for each factor k , and (c) for a model with K latent factors that satisfies conditions (a) and (b), the smallest number of items needed is the smallest integer J that satisfies

$$\begin{aligned} (J+1) + J(J+1)/2 \\ \geq (J+1) + (J+1-K)K + K + K(K+1) + 1, \end{aligned} \quad (6)$$

which means that the integer solution for smallest J is $J = 2K + 2$.

Condition (a) is equivalent to setting the variance of a latent factor to 1, because the latent factor does not have its own scale. This condition is often imposed for other latent factor models. Without imposing this condition, we can have equivalent models by dividing the

loadings for factor k and multiplying the standard deviation σ_{Zk} by a constant c . Thus condition (a) is necessary for model identification.

Condition (b) is imposed to avoid the rotational nonidentification illustrated in Section 3.2.1 by presetting some S_{jk} 's and their corresponding β_{jk} 's to 0. Matrix $\mathbf{B}$, satisfying conditions (a) and (b), can be written as

$$\begin{bmatrix} \beta_{11} & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & \beta_{KK} \\ \beta_{K+1,1} & \beta_{K+1,2} & \cdots & \beta_{K+1,K} \\ \vdots & \vdots & \vdots & \vdots \\ \beta_{J1} & \beta_{J2} & \cdots & \beta_{JK} \\ \gamma_1 & \gamma_2 & \cdots & \gamma_K \end{bmatrix},$$

with at least one β_{jk} in each column being 1. The $\mathbf{T}$ matrix that ensures the submatrix consisting of the first K rows of $\mathbf{B}^* (= \mathbf{B}\mathbf{T})$ diagonal and at least one β_{jk} in each column of $\mathbf{B}^*$ being 1 can only be the identity matrix. Thus $\{\mathbf{B}, \mathbf{\Sigma}, \mathbf{\alpha}, \mathbf{\Theta}\}$ cannot be rotated by $\mathbf{T}$ with condition (b) imposed. If any of the off-diagonal elements in the submatrix is nonzero, then $\mathbf{T}$ can be nonidentity matrix. Therefore, condition (b) is also necessary for model identification.

The condition (c) is the result of applying the t -rule postulated by Bollen (1989, p. 242); that is, the number of identifying equations cannot be smaller than the number of unknown parameters. The total number of identifying equations in (5) is specified on the left side of (6); there are $(J+1)$ marginal probabilities $\{P(U_{i1}=1), \dots, P(U_{iJ}=1), P(Y_i=1)\}$ plus $J(J+1)/2$ off-diagonal covariance elements between $[\text{logit}\mu_{i1}, \dots, \text{logit}\mu_{iJ}, \text{logit}(EY_i)]$. The diagonal elements of the covariance do not provide additional identifying equations, because they are the variances contained in the means in form of the probabilities. The maximum number of unknown parameters is specified in the right side of (6) for a model satisfying conditions (a) and (b), because there are $(J+1)$ elements in $\mathbf{B}_0 = (\beta_{01}, \dots, \beta_{0J}, \gamma_0)^T$, $(J+1-K)K$ elements in $\mathbf{B}$, K elements in $\mathbf{\alpha}$, $K(K+1)$ elements in $\mathbf{\Sigma}$ and $\mathbf{\Theta}$ combined, and one additional parameter γ_Z . For example, the maximum number of items required is 4 for a one-factor MSIRR model and 6 for a two-factor MSIRR model. For our gene methylation data, the maximum number of latent factors allowed is 5, which requires at least 12 genes (because we have a total of 13 genes).

When some additional S_{jk} 's are set to 0 by the specification search described in Section 4.2, conditions (a) and (b) are still needed, but the minimum J can be smaller because the number of unknown parameters is smaller. But because we do not know which additional S_{jk} 's are 0 before fitting a model, the number of gene items needs to satisfy condition (c).

For M2PL models, similar arguments for identification conditions can be established and we do not give them here due to the focus of this article and space limitations.

4. MODEL ESTIMATION AND SPECIFICATION SEARCHES

4.1 Estimation

Joint estimation of parameters in factor-structured models for categorical data is usually carried out by Markov chain Monte Carlo (MCMC) methods (Scott and Ip 2002; Wollack, Bolt, Cohen, and Lee 2002; Bolt and Lall 2003; Dunson 2003), maximum likelihood through the EM algorithm (Wollack et al. 2002; Wang, Chen, and Cheng 2004), or a Gaussian quadrature

numerical method (Hedeker and Gibbons 1994, Rabe-Hesketh, Skrondal, and Pickles 2002). Recent computational advances are crucial to successful estimation in relatively complex nonlinear models with multidimensional latent variables (Liu and Pierce 1994; Pinheiro and Bates 1995; McCulloch 1997; Clarkson and Zhan 2002; Rabe-Hesketh et al. 2002). For MCMC, a multivariate posterior distribution of the model parameters is the focus of inferential interest. In EM algorithm, latent variables are treated as missing data, and their conditional expectations are calculated in an E-step. Numerical methods based on Gaussian quadrature calculate the integrals through summing over quadrature points.

In our case, when a fixed set of selection indicators is given in a MSIRR model, joint estimation of the parameters through maximum likelihood can follow that of generalized linear, latent, and mixed models (GLLAMMs) (Rabe-Hesketh et al. 2000) or structural equation models with latent variables (Muthén and Muthén 2001). Let $\mathbf{S}=(\mathbf{s}_1^T, \dots, \mathbf{s}_J^T)^T$, where $\mathbf{s}_j=(S_{j1}, \dots, S_{jK})^T$, then the log-likelihood of our MSIRR model can be written as follows for a given selection matrix $\mathbf{S}$:

$$\sum_{i=1}^n \log \int_0^\infty \prod_{j=1}^J \frac{\exp\{U_{ij}(\beta_{0j} + \sum_{k=1}^K \beta_{jk} \eta_{ik} S_{jk})\}}{1 + \exp\{\beta_{0j} + \sum_{k=1}^K \beta_{jk} \eta_{ik} S_{jk}\}} \times \frac{\exp\{Y_i(\gamma_0 + \sum_{k=1}^K \gamma_k \eta_{ik} + \gamma_Z Z_i)\}}{1 + \exp\{\gamma_0 + \sum_{k=1}^K \gamma_k \eta_{ik} + \gamma_Z Z_i\}} g(\boldsymbol{\eta}_i) d\boldsymbol{\eta}_i, \quad (7)$$

where $\boldsymbol{\eta}_i = (\eta_{i1}, \dots, \eta_{iK})^T$ and $g(\boldsymbol{\eta}_i)$ denotes multivariate normal distribution of $\boldsymbol{\eta}_i$ with both mean and variance-covariance matrix depending on Z_i .

For the foregoing likelihood function, neither the expectation or the maximization can be computed in closed form in EM. Both Monte Carlo EM and MCMC rely heavily on numerical methods; thus we use Gaussian quadrature with 30 points in each dimension of the K latent factors to numerically maximize the log-likelihood given in (7). Standard errors of the parameter estimates are obtained by numerically inverting the negative Hessian matrix.

Purpose-written code in SAS version 9.1.2 that invokes proc nlmixed is used to fit the foregoing models. Our MSIRR can also be fitted using STATA module GLLAMM (Rabe-Hesketh et al. 2004). Both proc nlmixed in SAS and GLLAMM in STATA allow the user to specify quite flexible models with latent variables. GLLAMM even allows hierarchical structure in latent variables, whereas proc nlmixed allows latent variables in only one level.

4.2 Specification Searches With a Genetic Algorithm

The number of possible configurations of a selection matrix $\mathbf{S}$ can be calculated. For $K = 1$, there is only one $\mathbf{S}$, because all of the S_{j1} 's must be 1 for all j 's. For $K \geq 2$, the number of possible $\mathbf{S}$ matrices that satisfy our model constraint $\sum_{k=1}^K S_{jk} \geq 1$ and the identification conditions is

$$K! \binom{J}{K} (2^K - 1)^{J-K}, \quad (8)$$

which is 27,634,932 for $K = 2$ and 484,727,527,284 for $K = 3$ if $J = 13$. The foregoing formula is derived as follows. We first choose K items (rows) to be factorially simple, which is

$\binom{J}{K}$, and the K factorially simple items have $K!$ permutations. For each row in the rest of the $J - K$ items, there are $2^K - 1$ different combinations of 0 and 1. (The all-0 row is excluded due to the constraint.)

Thus examining all of the model configurations is an impossible task in real time for $K \geq 2$. The modification of an initially specified model to improve fit according to an objective function has been termed a *specification search* (Marcoulides and Drezner 2001). We conduct specification searches on the selection matrix $\mathbf{S}$ of dimension $J \times K$ with a GA to discover our data structure. Adaptive search procedures like GAs have proven quite effective in such large-scale optimization problems.

A GA is one type of adaptive search procedure that performs a multidimensional search so that several solutions can be considered simultaneously by maintaining a population of potential solutions. The population of potential solutions undergoes a simulated evolution in which the relatively good solutions reproduce at each generation while relatively bad solutions die. The striking characteristic of a GA is that model parameters are not directly manipulated, but rather a coding of the parameter set (in our case the $\mathbf{S}$ selection matrix) is directly manipulated. (For a detailed introduction to GAs, see Goldberg 1989 and references cited therein.) In the remainder of this section, we describe how a GA is implemented in our particular case.

We use a bias-corrected Akaike information criterion (AICC) as an objective function for model selection. The AICC is a second-order variant of the Akaike information criterion (AIC) that adjusts small-sample bias (Hurvich and Tsai 1989). If the sample size is large with respect to the number of parameters, then the second-order correction is negligible, and the AIC and AICC will tend to select the same model. When the sample size is relatively small, the AICC is preferred, because it penalizes more of the complexity in a model by correcting bias in small samples. Burnham Anderson, and White (1994) and Janssen and Boeck (1999) found the AICC approximation to the Kullback–Leibler distance to be especially useful in product multinomial models and multidimensional item response models. Another point that we want to emphasize is that although two models that differ in $\mathbf{S}$ may be nested, we do not want to use nonpenalized likelihood ratio test to decide which is the better model, because there might be considerable structural changes in the latent factors when $\mathbf{S}$ is changed. In addition, we require that the candidate models in our case must satisfy $\gamma_k > 0$ —in (3) for all k 's, because the objective is to identify SV40 altered functional groups that increase the probability of a cancer case. We posit the restriction $\gamma_k \geq 0$ —instead of $\gamma_k > 0$, so that $\gamma_k = 0$ is legitimate parameter value for the likelihood ratio test for a chosen model according to the AICC.

To generate an initial selection matrix for $K \geq 2$, we start with a pre- $\mathbf{S}$ matrix of 0's with J rows and K columns. We first randomly select K rows in the matrix and set only one element of different columns in the K rows to 1, so that the second identification condition is satisfied. For the rest of the $J - K$ rows in the pre- $\mathbf{S}$ matrix, we randomly assign 1 to elements of the first $K - 1$ columns in each row. If all of the $K - 1$ elements in a row are 0, then the K th element in the row is set to 1; otherwise, the K th element is randomly assigned to 1 or 0, so that

$\sum_{k=1}^K S_{jk} \geq 1$. No objective function can be used to evaluate models with different response vectors if $\sum_{k=1}^K S_{jk} = 1$, although they can be used to evaluate models with different predictor variables. (In our model, the only predictor variable is Z .)

We then arrange the elements in a $\mathbf{S}$ matrix by column to form a $J \times K$ vector that is called a chromosome. For example, an initial chromosome for the 13 genes with $K = 3$ could be coded as 1001111xxxxx010xxx111xxx001xxxxxx111, where x may be either 0 or 1, with the order within the 13 genes as shown in Table 1. Some initial chromosomes (starting values) are

generated, the best six in the AICC with $\gamma_k > 0$ are selected as our starting population, and the population size is fixed throughout generations. It should be noted that the larger the number of chromosomes in a population, the smaller the number of generations needed to find an optimal model, but the longer the computing time needed in each generation. Selection, mutation, and crossover (recombination) operators are applied on parent chromosomes to produce a new generation of chromosomes. Briefly, a chromosome is selected for mutation, and a pair of chromosomes is selected for crossover. Mutation means that a binary code is changed (e.g., a “0” may be mutated to “1” and vice versa). Crossover means that two chromosomes are recombined at a certain breakpoint of the chromosomes to form a new chromosome. Each resulting chromosome is checked to see whether the second identification condition and $\sum_{k=1}^K S_{jk} \geq 1$ are satisfied; if not, then proper mutation is reapplied to ensure the conditions. An offspring chromosome is retained if it is among the six best-fitted chromosomes in the new generation according to the foregoing selection criteria; otherwise, it is left to die.

We apply a descent algorithm on the offspring until no better solution can be found by changing a single code on a chromosome. The GA is stopped when the population does not change for three consecutive generations. Although it was not necessary, we repeated the GA several times and found rather consistent results.

The first identification condition is ensured later in the parameter estimation stage by setting one of the factor loadings to 1 for each k once the configuration of $\mathbf{S}$ matrix is determined by the GA. The third identification condition is always satisfied in our case as long as K does not exceed 5.

5. ANALYSIS OF THE GENE METHYLATION DATA

5.1 Fitting the MSIRR to the Gene Methylation Data

For $K = 1$, we do not need to use the GA, because all of the elements of $\mathbf{S}$ are 1. We now describe how the GA is applied for $K = 2$. Each chromosome is thus a vector of 26 binary codes, and we choose the population size as 6. When crossover is applied in two parent chromosomes, the recombined chromosome sometimes must go through a mutation to produce an offspring

in which the constraint $\sum_{k=1}^K S_{jk} \geq 1$ and the second identification condition postulated in Section 3.2.2 are satisfied. The third identification condition in our case is satisfied, because $J = 13$ is a sufficiently large value for K up to 5. Table 2 shows the evolution of the last eight generations of the chromosome that resulted in the final solution for $K = 2$, but not that the other chromosomes. The chromosome in generation -3 is the result of recombination of the two chromosomes from the previous generation; it can be seen that the AICC value of the recombined chromosome is greatly reduced from those of its parents (i.e., the smaller the AICC, the better the fit). The second and third elements in the second column of $\mathbf{S}_1$ (corresponding to the 15th and 16th binary codes in the chromosome) in generation -3 were mutated, which resulted a further reduced AICC; mutations in other elements did not improve the AICC by much. The two elements are selection indicators for *DcR1* and *DcR2*, and we found the factor loadings for these two genes in generation -3 were negative, although nonsignificant, at .05. These findings suggest that mutations from 1 to 0 on them likely would improve the AICC, and the chromosome in generation -2 confirmed the conjecture. The decent algorithm was applied at the offspring of the generation -2, a single gene mutation at *RASSF1* of the second pathway, changing S_{92} from 0 to 1 resulted in further improvement of the AICC.

The final selection matrix (generation 0 in Table 2) that we created with has the first column as all 1's except the selection indicator for *RASSF1*, which is the only factorially simple item in the first functional group in the methylation pathway. Aberrant methylation of *DAP-K*,

P15, *P73*, *RASSF1*, *TIMP3*, and *SHP1* is involved in the second functional group of the pathway.

For $K = 3$, the GA is applied in a similar fashion. However, at least one of the γ_k 's is negative in all three functional group models that we examined, even though the α_k 's are all positive. This indicates that solutions with $K = 3$ are not supported by our data. Thus we identified two functional groups of the methylation pathway from SV40 to the cancer for our data.

5.2 Results From Analysis of the Gene Methylation Data

For $K = 1$, all of the β_{j1} 's are highly significant, indicating that all of the genes were involved in the methylation pathway. The α_1 's are also highly significant, indicating that SV40 increased the level of aberrant methylation. Aberrant methylation significantly increases the likelihood of being a case, because γ_1 is highly significant. The γ_Z is not significant, indicating that other SV40-nonmethylation oncogenic pathways do not seem to exist in our samples.

For $K = 2$, Figure 2 presents a graphical diagram of the optimal links between the genes and the two identified functional groups of aberrant methylation. Once a final model is obtained, we use the Wald test to assess the effects of statistical interest. We point out that whether a coefficient is significant or not at a type I error of .05 depends on the sample size and on the precision of the measurement; this is also true for other types of models in which Wald tests are administrated.

The value of a selection indicator S_{jk} is determined by the GA and indicates the relative importance of the gene j in functional group k of the methylation. $S_{ij} = 1$ indicates that gene j is important to functional group k . The statistical significance of coefficient β_{jk} indicates that the extent of aberrant methylation of gene j is regulated by the functional group k and that the actual value of β_{jk} is the effect size of the functional group k having an aberrant methylation of gene j relative to that of the gene with a factor loading set to 1 for factor k [see identification condition (a) in Sec. 3.2.2]. Table 3 shows that all genes belong to the first functional group of aberrant methylation except *RASSF1*, which was chosen by the GA as the only factorially simple item for a possible second functional group in our final model. As judged by the Wald statistics, the coefficients of factor loadings β_{j1} are all significant at .05 except those of *SHP1* and *CDH13*. The factor loading β_{11} for *DAP-K* in the first group is significant, because we can assess its significance by switching the labels for genes 1 and 12 (or any other genes), and we subsequently can arrive at the same conclusion by examining the ratios of the $\sigma_{z_1}^2$'s to their corresponding standard errors. Note that *SHP1* is aberrantly methylated in almost all of the cancer cases, resulting in a large standard error estimate for its $\beta_{0,11}$ and $\beta_{11,k}$. Only *DAP-K*, *P15*, *P73*, *RASSF1*, *TIMP3*, and *SHP1* are involved in the second functional group of aberrant methylation. The coefficients β_{j2} are significant at .05 for all of these genes except *SHP1*. β_{92} for *RASSF1* is significant for the one-tailed test of alternative hypothesis $\beta_{92} \geq 0$ but only marginally significant for the two-tailed test.

It can be seen that the outcome of leukemia/lymphoma is positively related to both identified functional groups of aberrant methylation, as suggested by the highly significant γ_1 and γ_2 . After accounting for the roles of the both functional groups of aberrant methylation, the association between SV40 and cancer is no longer detectable, as suggested by the nonsignificant γ_Z (sixth row from the bottom in Table 3). This confirms that the aberrant methylation of the tumor-suppressor genes lies between SV40 and the cancer and also indicates that pathways from SV40 to cancer other than aberrant methylation are unlikely. SV40 significantly boosts the aberrant methylation, as suggested by the highly significant α_k 's, especially in the first functional group. We found that the variances of both latent factors are much larger in noninfected samples than in SV40 infected samples (third and fourth rows from

bottom in Table 3), suggesting that the level of aberrant gene methylation is relatively more homogeneously elevated in SV40-detectable samples. There is no correlation between the aberrant methylation levels of the first and second functional groups (the last two rows in Table 3), which means the two identified groups have quite distinct functions in methylation, even though common genes are involved in both groups.

We thus identified a general functional group of genes in aberrant methylation comprising all of the 13 genes and the other functional group consisting of cell cycleincluding apoptosis-regulatory genes (*P15*, *P73*, *RASSF1A*, and *DAPK*), the gene in the Jak/Stat signaling pathway (*SHP1*), and the gene of tissue inhibitor of metalloproteinase 3 (*TIMP3*).

Because the current literature on gene methylation in hematologic cancer has not indicated grouping or co-regulation of genes, the findings from our MSIRR models will provide guidance for further scientific investigation.

5.3 Comparison With Analysis From the “Full” Factor Structured Model

A model in which all of the genes are factorially complex—that is, a model in which all of the genes are linked to all of the latent factors [such as an exploratory-type linear factor or item response model satisfying condition (a) but with unknown covariance for latent factors]—is not identifiable, although most published research articles have adopted such a model specification by using only certain rotations or assuming that the covariance matrix of the latent factors are known to be diagonal (Hojtink and Rooks 1999; Wang et al. 2004; Moustaki et al. 2004). Additional identification conditions, such as condition (b), must be imposed on the full factor structured model for binary response for the sake of identification. Furthermore, without the aid of the genetic search algorithm, it would not be feasible at best or possible in practice to make connections between the genes and a specific functional group of aberrant methylation due to the large number of possible connections, as shown in (8). Therefore, here we do not fit the conventional full exploratory factor model, but rather fit a “full” factor structured model with 2 latent factors in which 2 of the 13 genes are factorially simple and all of the rest are linked to the 2 factors.

We now examine the result from a two-dimensional “full” factor structured model in which *RASSF1A* and *RARB* are factorially simple for the two factors (Table 4). Note that there are

total of $2! \binom{13}{2} = 156$ different “full” factor structured models for $K = 2$, and we pick *RASSF1A* and *RARB* to be factorially simple out of the 156 possible choices, to be in line with the result from the forgoing MSIRR model. Note that without our GA search, we would have not known to let *RASSF1A* and *RARB* be factorially simple and thus would not have gotten such similar results from the “full” factor structured model (shown in Table 4) and our MSIRR model. The inference from the first functional group is identical to that of the final model presented in Table 3. For the second group, along with *DAP-K*, *P15*, *P73*, *RASSF1*, *TIMP3*, and *SHP1*, which are common in both “full” factor structured model and the model in Table 3, all the remaining genes except *RARB* are also present in this group. It can be seen that the inference for *DAP-K*, *P73*, *RASSF1*, *TIMP3*, and *SHP1* in the second group is identical to that of the final model in Table 3; here the factor loadings β_{62} for *P15* are significant at .05 for the one-tailed test and marginally significant for the two-tailed test, whereas they are significant for the two-tailed test in Table 3. None of the genes in the second functional group that are not detected in the final model of Table 3 has a significant factor loading in the “full” factor structured model, and actually the factor loadings for *DcR1*, *DcR2*, and *CRBP* are negative.

We stress that the “full” factor structured model here differs from the usual full exploratory model in that the identification condition (b) is imposed on the former but not on the latter. A usual full exploratory model requires no factorially simple items and thus is not identifiable

unless the covariance of the factors is not estimated freely. In the “full” factor structured model that we fit, we arbitrarily let *RASSF1* and *RARB* be the two factorially simple items. This model is the aftermath one we chose that closely matches our final optimal model derived through the GA in Table 3. Note that there is no guarantee that a “matched” “full” factor structured model and the corresponding optimal MSIRR model will give similar results, because, as mentioned earlier, if we were to drop a nonsignificant term in model (1) from a larger model by setting additional S_{jk} 's to 0, then the structure of the latent factors would change, and thus the results would differ. This point is illustrated by the finding that not all of the β 's in the identified functional groups from our final MSIRR model were statistically significant, but removing the corresponding gene items with nonsignificant β 's resulted in more dramatic changes in significance for α_k and α_k (results not shown). Thus a “matched” “full” factor model can have a different set of statistically significant β_{jk} 's than a final MSIRR model, in which case the “matched” model could not be an optimal model, because the GA search space should have covered this “matched” model.

6. DISCUSSION

We have presented a multidimensional selective item response regression model. The model is multidimensional because we specify multiple correlated latent factors; it is item-response in nature because we relate items to underlying latent factors; it is selective because items are selected to be linked to a latent factor, and it is a latent regression model, because the distribution of the latent factors (both means and covariance matrix) depends on input (here the SV40), and the outcome is also regressed on input and the latent factors. We echo the opinion of Rabe-Hesketh and Skrondal (2001) that the factor-structured model should be more widely used outside the fields of psychological measurements and educational testing for which it was originally developed, and that fundamental statistical issues of identification should be given more attention.

We showed how the GA can be applied to guide our discovery of the data structure. Although the GA is an automatic algorithm, inspection of statistical significance of the model parameters and scientific information may guide the selection of both mutation sites and recombination points, possibly decreasing the number of generations needed to arrive at an optimal model. Like other nonlinear models with latent variables, equivalent MSIRR models in terms of the AICC with different configurations in **S** may be encountered with GA specification searches; if this occurs, it is the responsibility of the subject matter experts to determine which model to accept as the best model. No automated specification search can make such a decision. Furthermore, we must admit that evaluating even one chromosome is computationally intensive with multiple binary responses and latent variables.

Our results from the MSIRR model show that the SV40- methylation pathway is crucial in hematopoietic malignancies, but we cannot exclude other non-SV40-related pathways in other data sets. Another scientific possibility is that SV40 sequences may directly enhance methylation in a tumor-suppressor gene beyond their effects through the functional groups. We further investigate this hypothesis by adding Z_i as a covariate to the final optimal model (1) (as selected by the GA in Table 2), so (1) becomes

$$\text{logit}\mu_{ij} = \beta_{0j} + \sum_{k=1}^K \beta_{jk} \eta_{ik} S_{jk} + \beta_{Z_j} Z_i.$$

We added covariate Z_i for each j except for two j 's that are factorially simple for the two identified functional groups of aberrant methylation, to ensure identification, and estimated the parameters through the maximum likelihood method. None of the coefficients β_{Z_j} 's was

significant (results not shown). This confirms that SV40 induces changes in the aberrant methylation level through the two functional groups, and the association between SV40 and the aberrant methylation can be fully explained by the identified SV40-methylation pathway.

In our study, we had no missing-item responses. If missing items are present, then the observed data log-likelihood in (7) can be modified to replace the J in the upper product limit by J_i , where $J_i \leq J$ are the observed item response variables in sample i . If the maximum likelihood method is used to jointly estimate the parameters given a selection matrix $\mathbf{S}$, then missing-item responses that are missing at random can be accommodated. For nonignorable missing items, additional modeling of the missing items must be done.

Acknowledgments

This work was supported in part by National Institute of Mental Health grant 1R01MH66187-01A2 and by National Cancer Institute Early Detection Research Network grants 5U01CA086368 and U01CA084971. The authors thank Ralitzia Gueorguieva for helpful discussions.

REFERENCES

- Bollen, KA. Structural Equations With Latent Variables. New York: Wiley; 1989.
- Bolt DM, Lall VF. Estimation of Compensatory and Noncompensatory Multidimensional Item Response Models Using Markov Chain Monte Carlo. *Applied Psychological Measurement* 2003;27:395–414.
- Burnham KP, Anderson DR, White GC. Evaluation of the Kullback-Leibler Discrepancy for Model Selection in Open Population Capture-Recapture Models. *Biometrical Journal* 1994;36:299–315.
- Clarkson DB, Zhan YH. Using Spherical-Radial Quadrature to Fit Generalized Linear Mixed, Effects Models. *Journal of Computational and Graphical Statistics* 2002;11:639–659.
- Coull BA, Agresti A. Random Effects Modeling of Multiple Binomial Responses Using the Multivariate Binomial Logit-Normal Distribution. *Biometrics* 2000;56:73–80. [PubMed: 10783779]
- Dammann R, Schagdarsurengin U, Strunnikova M, Rastetter M, Seidel C, Liu L, Tommasi S, Pfeifer GP. Epigenetic Inactivation of the RAS-Association Domain Family 1 (*RASSF1A*) Gene and Its Function in Human Carcinogenesis. *Histology and Histopathology* 2003;18:665–677. [PubMed: 12647816]
- Dunson DB. Bayesian Latent Variable Models for Clustered Mixed Outcomes. *Journal of the Royal Statistical Society Ser. B* 2000;62:355–366.
- Dunson DB. Dynamic Latent Trait Models for Multidimensional Longitudinal Data. *Journal of the American Statistical Association* 2003;98:555–563.
- Fisher SG, Weber L, Carbone M. Cancer Risk Associated With Simian Virus 40–Contaminated Polio Vaccine. *Anticancer Research* 1999;19:2173–2180. [PubMed: 10472327]
- Goldberg, DE. Genetic Algorithms in Search, Optimization, and Machine Learning. Reading, MA: Addison-Wesley; 1989.
- Hedeker D, Gibbons RD. A Random-Effects Ordinal Regression Model for Multilevel Analysis. *Biometrics* 1994;50:933–944. [PubMed: 7787006]
- Hojtink H, Rooks G. Confirmatory Factor Analysis of Items With a Dichotomous Response Format Using the Multidimensional Rasch Model. *Psychological Methods* 1999;3:330–314.
- Hurvich CM, Tsai C-L. Regression and Time Series Model Selection in Small Samples. *Biometrika* 1989;76:297–307.
- Ihara M, Kano Y. Identifiability of Full, Marginal, and Conditional Factor Analysis Models. *Statistics & Probability Letters* 1995;23:343–350.
- Janssen R, Boeck PD. Confirmatory Analyses of Componential Test Structure Using Multidimensional Item Response Theory. *Multivariate Behavioral Research* 1999;34:245–268.
- Jones PA, Laird PW. Cancer Epigenetics Comes of Age. *Nature Genetics* 1999;21:163–167. [PubMed: 9988266]
- Jöreskog, KG.; Sörbom, D. LISREL 8 User's Reference Guide. Chicago: Scientific Software International; 1999.
- Liu Q, Pierce DA. A Note on Gauss-Hermite Quadrature. *Biometrika* 1994;81:624–629.

- Marcoulides, GA.; Drezner, Z. Specification Searches in Structural Equation Modeling With a Genetic Algorithm. In: Marcoulides, GA.; Schumacker, RE., editors. *New Developments and Techniques in Structural Equation Modeling*. Mahwah, NJ: Lawrence Erlbaum; 2001. p. 247-268.
- McCulloch CE. Maximum Likelihood Algorithms for Generalized Linear Mixed Models. *Journal of the American Statistical Association* 1997;92:162-170.
- Moustaki I, Jöreskog KG, Mavridis D. Factor Models for Ordinal Variables With Covariate Effects on the Manifest Variables and Latent Variables: A Comparison of LISREL and IRT Approaches. *Structural Equation Modeling* 2004;11:487-513.
- Muthén, BO.; Muthén, LK. *Mplus: Statistical Analysis With Latent Variables*. Los Angeles: Muthén & Muthén; 2001.
- Paulson EJ, Fingerroth JD, Yates JL, Speck SH. Methylation of the EBV Genome and Establishment of Restricted Latency in Low-Passage EBV-Infected 293 Epithelial Cells. *Virology* 2002;299:183-214.
- Pinheiro JC, Bates DM. Approximations to the Log-Likelihood Function in the Nonlinear Mixed-Effects Model. *Journal of Computational and Graphical Statistics* 1995;4:12-35.
- Rabe-Hesketh S, Skrondal A. Parameterization of Multivariate Random Effects Models for Categorical Data. *Biometrics* 2001;57:1256-1263. [PubMed: 11764269]
- Rabe-Hesketh S, Skrondal A, Pickles A. Reliable Estimation of Generalized Linear Mixed Models Using Adaptive Quadrature. *The Stata Journal* 2002;2:1-21.
- Rabe-Hesketh, S.; Skrondal, A.; Pickles, A. *GLLAMM Manual*. Berkeley, CA: Be Press; 2004.
- Rabe-Hesketh S, Skrondal A, Taylor C. *Generalized Linear Latent and Mixed Models*. *Stata Technical Bulletin* 2000;53:47-57.
- Reckase, MD. A Linear Logistic Multidimensional Model for Dichotomous Item Response Data. In: Linden, WJ.; Jambleton, RK., editors. *Handbook of Modern Item Response Theory*. New York: Springer; 1997. p. 271-286.
- Scott SL, Ip EJ. Empirical Bayes and Item-Clustering Effects in a Latent Variable Hierarchical Model: A Case Study From the National Assessment of Educational Progress. *Journal of the American Statistical Association* 2002;97:409-419.
- Shivapurkar N, Harada K, Reddy J, Scheuermann RH, Xu Y, McKenna RW, Milchgrub S, Kroft SH, Feng Z, Gazdar AF. Presence of Simian Virus 40 DNA Sequences in Human Lymphomas. *Lancet* 2002;359:851-852. [PubMed: 11897287]
- Shivapurkar N, Takahashi T, Reddy J, Zheng Y, Stastny V, Collins R, Toyooka S, Suzuki M, Parikh G, Asplund S, Kroft SH, Timmons C, McKenna RW, Feng Z, Gazdar AF. Presence of Simian Virus 40 DNA Sequences in Human Lymphoid and Hematopoietic Malignancies and Their Relationship to Aberrant Promoter Methylation of Multiple Genes. *Cancer Research* 2004;64:3757-3760. [PubMed: 15172980]
- Toyooka S, Pass HI, Shivapurkar N, Fukuyama Y, Maruyama R, Toyooka KO, Gilcrease M, Farinas A, Minna JD, Gazdar AF. Aberrant Methylation and Simian Virus 40 Tag Sequences in Malignant Mesothelioma. *Cancer Research* 2001;61:5727-5730. [PubMed: 11479207]
- Wang WC, Chen PH, Cheng YY. Improving Measurement Precision of Test Batteries Using Multidimensional Item Response Models. *Psychological Methods* 2004;9:116-136. [PubMed: 15053722]
- Wollack JA, Bolt DM, Cohen AS, Lee Y-S. Recovery of Item Parameters in the Nominal Response Model: A Comparison of Marginal Maximum Likelihood Estimation and Markov Chain Monte Carlo Estimation. *Applied Psychological Measurement* 2002;26:339-352.
- Wong TS, Man MWL, Lam AKY, Wei WI, Kwong YL, Yuen APW. The Study of p16 and p15 Gene Methylation in Head and Neck Squamous Cell Carcinoma and Their Quantitative Evaluation in Plasma by Real-Time PCR. *European Journal of Cancer* 2003;39:1881-1887. [PubMed: 12932666]
- Zhang YJ, Chen Y, Ahsan H, Lunn RM, Lee PH, Chen CJ, Santella RM. Inactivation of the DNA Repair Gene O-6-MGMT by Promoter Hypermethylation and Its Relationship to Aflatoxin B-1-DNA Adducts and p53 Mutation in Hepatocellular Carcinoma. *International Journal of Cancer* 2003;103:440-444.

Figure 1.
Graphical representation of the relationship among SV40, genes, and cancer.

Figure 2.
Graphical representation of the two-group solution.

Table 1
Correlation matrix for SV40, aberrant gene methylation, and leukemia/lymphoma

	DAP-K	DcR1	DcR2	CRBP	P16	P15	CDH1	RARB	RASSF1	TIMP3	SHP1	P73	CDH13	SV40	Cancer
DAP-K	1														
DcR1	+	1													+
DcR2	+	+	1												-
CRBP	+	+	+	1											+
P16	+	+	+	+	1										-
P15	+	+	+	+	+	1									+
CDH1	+	+	+	+	+	+	1								+
RARB	+	+	+	+	+	+	+	1							+
RASSF1	+	+	+	±	-	-	-	-	1						-
TIMP3	+	±	+	+	-	-	+	+	+	1					+
SHP1	+	±	+	+	-	-	+	+	+	+	1				+
P73	+	+	-	-	+	+	+	+	+	+	+	1			+
CDH13	+	+	+	+	+	+	+	+	±	+	+	+	1		+
SV40	+	+	+	+	+	+	+	+	±	+	+	±	+	1	+
Cancer	+	-	+	-	-	+	+	+	-	+	+	+	+	+	1

NOTE: +, statistically significantly correlated with $p < .05$; ±, marginally correlated with $p > .05-.08$; -, not significantly correlated with $p > .08$.

Table 2
Evolution of a chromosome in the last eight generations

No.	Gene	Generation index									
		-4		-3		-2		-1		0	
		S ₁	S ₂	S ₁ ^r	S ₁	S ₁	S ₁	S ₁	S ₁	S ₁	S ₁
1	DAP-K	11	11	11	11	11	11	11	11	11	11
2	DcR1	01	10	11	11	10	10	10	10	10	11
3	DcR2	01	10	11	10 ^m	10 ^m	10 ^m	10	10	10	10 ^m
4	CRBP	10	10	10	10	10	10	10	10	10	10 ^m
5	P16	10	11	10	10	10	10	10	10	10	10
6	P15	01	01	01	01	01	01	01	1 ^m 1	10	11
7	CDH1	10	10	10	10	10	10	10	11 ^m	11	11
8	RARB	10	10	10	10	10	10	10	10	10	11
9	RASSF1	10	10	10	10	10	11 ^d	0 ^m 1	01	01	01
10	TIMP3	10	10	10	10	10	10	11 ^m	11	01	01
11	SHP1	11	11	11	11	11	11	11	11	11	1 ^m 1
12	P73	01	01	01	01	01	01	01	1 ^m 1	11	11
13	CDH13	10	10	10	10	10	10	10	10	11 ^m	11
AICC		2,585.8	2,532.8	2,509.7	2,507.8	2,503.5	2,498.3	2,481.2	2,496.2	2,485.6	2,513.8

^rCrossover was applied between the two parent chromosomes $\text{vec}(\mathbf{S}_1)$ and $\text{vec}(\mathbf{S}_2)$ at the end of the 13th element. The offspring consists of the first column of $\mathbf{S}_2$ and the second column of $\mathbf{S}_1$ from the previous generation.

^{rr}Crossover was applied between the two parent chromosomes $\text{vec}(\mathbf{S}_1)$ and $\text{vec}(\mathbf{S}_2)$ at the end of the 13th element. The offspring consists of the second column of $\mathbf{S}_2$ and the second column of $\mathbf{S}_1$ from the previous generation.

^dA decent algorithm was applied on the previous offspring.

^mMutation was applied on the previous generation. Generation 3 is derived from $\mathbf{S}_2$ of generation 2, because $\mathbf{S}_1$ in generation 2 does not survive to generation 3.

Table 3

Parameter estimates in the two-group solution (standard errors in parentheses)

No.	Gene	β_{0j}	β_{j1} ($k = 1$), first group	β_{j2} ($k = 2$), second group
1	<i>DAP-K</i>	-2.43 _(.44)	1.00 _(.0)	1.00 _(.00)
2	<i>DcR1</i>	-2.20 _(.40)	.96 _(.27)	
3	<i>DcR2</i>	-3.19 _(.52)	.92 _(.28)	
4	<i>CRBP</i>	-1.36 _(.24)	.60 _(.18)	
5	<i>P16</i>	-2.16 _(.32)	.62 _(.19)	
6	<i>P15</i>	-1.49 _(.23)	.40 _(.13)	.57 _(.29)
7	<i>CDH1</i>	-1.21 _(.31)	1.02 _(.28)	
8	<i>RARB</i>	-1.87 _(.29)	.64 _(.19)	
9	<i>RASSF1</i>	-3.82 _(.76)		1.54 _(.83)
10	<i>TIMP3</i>	-2.90 _(.45)	.78 _(.23)	.82 _(.41)
11	<i>SHP1</i>	-61.26 _(121.74)	13.44 _(32.66)	229.11 _(437.02)
12	<i>P73</i>	-4.24 _(.89)	.43 _(.21)	1.89 _(.91)
13	<i>CDH13</i>	-220.38 _(317.02)	125.37 _(184.28)	
	γ	-.24 _(1.54)	1.03 _(.41)	8.51 _(3.44)
	α_k		2.77 _(.60)	.44 _(.18)
	σ_{0k}^2		2.06 _(.46)	.98 _(.32)
	σ_{1k}^2		1.17 _(.33)	.67 _(.31)
	ρ_0	.0045 _(.0083)		
	ρ_1	.039 _(.059)		

NOTE: $-2 \times \log\text{-likelihood} = 2,378.2$; AICC = 2,481.2.

Table 4

Parameter estimates in the two-group solution of the nonselective model (standard errors in parentheses)

Gene	β_{0j}	β_{j1} ($k = 1$), first group	β_{j2} ($k = 2$), second group
<i>DAP-K</i>	-2.38 _(.37)	1.00 _(.0)	1.00 _(.0)
<i>DcR1</i>	-2.37 _(.46)	1.12 _(.28)	-.76 _(.45)
<i>DcR2</i>	-3.32 _(.57)	1.02 _(.28)	-.50 _(.36)
<i>CRBP</i>	-1.39 _(.26)	.67 _(.17)	-.47 _(.29)
<i>P16</i>	-2.32 _(.38)	.74 _(.20)	.90 _(.60)
<i>P15</i>	-1.48 _(.23)	.42 _(.13)	.56 _(.30)
<i>CDH1</i>	-1.21 _(.30)	1.03 _(.22)	.03 _(.34)
<i>RARB</i>	-1.87 _(.29)	.64 _(.16)	
<i>RASSF1</i>	-3.55 _(.60)		1.46 _(.77)
<i>TIMP3</i>	-2.94 _(.46)	.83 _(.22)	.82 _(.40)
<i>SHP1</i>	-53.15 _(90.53)	14.47 _(26.69)	223.11 _(424.18)
<i>P73</i>	-4.04 _(.84)	.53 _(.24)	2.03 _(.57)
<i>CDH13</i>	201.40 _(10.36)	116.6 _(173.15)	.02 _(1.92)
γ	-.07 _(1.45)	1.10 _(.40)	9.55 _(2.34)
α_k		2.91 _(.43)	.33 _(.10)
σ_{0k}^2		2.02 _(.32)	.87 _(.17)
σ_{1k}^2		1.26 _(.34)	.51 _(.21)
ρ_0	.0025 _(.12)		
ρ_1	-.055 _(.039)		

NOTE: $-2 \times \log\text{-likelihood} = 2,373.6$; AICC = 2,496.2.