Model-Based Recursive Partitioning of Extended Redundancy Analysis with an Application to Nicotine Dependence among US adults

Sunmee Kim[1] and Heungsun Hwang[*2]

[1] University of Manitoba, Winnipeg, Canada

[2] McGill University, Montreal, Canada

*Corresponding author information: Heungsun Hwang, Department of Psychology, McGill University, 2001 McGill College Avenue, Montreal, QC H3A 1G1, Canada. Email: heungsun.hwang@mcgill.ca.

This is a post-peer-review, pre-copyedit version of an article accepted for publication in British Journal of Mathematical and Statistical Psychology. Please do not copy or cite without authors' permission. The final article will be available, upon publication, via its DOI: 10.1111/bmsp.12240

Abstract

Extended redundancy analysis (ERA) is used to reduce multiple sets of predictors to a smaller number of components and examine the effects of these components on a response variable. In various social and behavioral studies, auxiliary covariates (e.g., gender, ethnicity, etc.) can often lead to heterogeneous subgroups of observations, each of which involves distinctive relationships between predictor and response variables. ERA is currently unable to consider such covariate-dependent heterogeneity to examine whether the model parameters vary across subgroups differentiated by covariates. To address this issue, we combine ERA with model-based recursive partitioning in a single framework. This combined method, MOB-ERA, aims to partition observations into heterogeneous subgroups recursively based on a set of covariates while fitting a specified ERA model to data. Upon the completion of the partitioning procedure, one can easily examine the difference in the estimated ERA parameters across covariate-dependent subgroups. Moreover, it produces a tree diagram that aids in visualizing a hierarchy of partitioning covariates, as well as interpreting their interactions. In the analysis of public data concerning nicotine dependence among US adults, the method uncovered heterogeneous subgroups characterized by several sociodemographic covariates, each of which yielded different directional relationships between three predictor sets and nicotine dependence.

Keywords: Extended redundancy analysis, model-based recursive partitioning, covariate-dependent heterogeneity, decision tree, model visualization

Introduction

Extended redundancy analysis (ERA; Takane & Hwang, 2005) is a statistical method that relates multiple sets of predictors to response variables. In ERA, a component is extracted from each set of predictors in such a way that it accounts for the maximum variation of a response variable. In this regard, ERA aims to perform data reduction and linear regression simultaneously, providing a simpler description of directional relationships among many sets of variables. ERA has been extended to improve its data-analytic flexibility, including generalized ERA for the analysis of a response variable that arises from an exponential-family distribution (Lee et al., 2016), functional ERA for the analysis of smooth functions or curves (Hwang et al., 2015; Tan, Choi, & Hwang, 2015), multivariate ERA for the analysis of multiple correlated responses (Kim et al., 2020; Lee et al., 2019; Lee et al., 2018), and Bayesian ERA (Choi, Kyung, Hwang, & Park, 2019).

In many social and behavioral studies, researchers often identify heterogeneous subgroups of observations based on auxiliary covariates, e.g., age, gender, ethnicity, etc., each of which involves different strengths/directions of relationships between variables of interest (Merkle & Zeileis, 2013; Raudenbush, 1997; Royston & Sauerbrei, 2004; Shadish, Cook, & Campbell, 2002). For example, many nicotine dependence studies show that the effects of occupation type, alcohol consumption pattern, or physical activity level on smoking initiation or cessation differ by ethnicity and race (Daza et al., 2006; Hu, Davies, & Kandel, 2006; Kandel, Kiros, Schaffran, & Hu, 2004; Robinson et al., 2006). In psychological and educational testing, item bias or differential item functioning is often present between different gender or cultural groups (Cauffman & MacIntosh, 2006; Fleishman, Spector, & Altman, 2002; Smith & Reise, 1998; Strobl, Kopf, & Zeileis, 2015). In pediatric obesity studies, the relationship between

obesity and its predictors related to impaired health-related quality of life is shown to vary across sex, race, and/or nations (Maher, 2004; Wake, Salmon, Waters, Wright, & Hesketh, 2002; Williams, Wake, Hesketh, Maher, & Waters, 2005; Zeller & Modi, 2006). Moreover, the growth rate of intelligence in early childhood appears to be divergent across parental SES groups (Brandmaier, von Oertzen, McArdle, & Lindenberger, 2013; McArdle & Epstein, 1987; Von Stumm & Plomin, 2015). Although such covariate-dependent heterogeneity is prevalent in practice, ERA has no mechanism to account for this heterogeneity efficiently, thus being unable to examine whether the relationships between predictor and response variables vary across subgroups of observations differentiated by additional covariates.

One may attempt to investigate covariate-dependent heterogeneity in ERA using a multiple-group analysis, where researchers prespecify relevant covariates (and subgroups derived from different combinations of the covariates) ahead of data analysis and examine differences in the ERA parameter values across the subgroups. In practice, however, it is difficult to know *a priori* which covariates (and their interactions) might affect the parameter heterogeneity. In addition, possible combinations of comparison subgroups are numerous when there are continuous covariates, categorical covariates with multiple levels, and/or a number of covariates at the same time (Strobl et al., 2015; Su, Tsai, Wang, Nickerson, & Li, 2009; Zeileis, Hothorn, & Hornik, 2008).

To address this issue, we propose to combine ERA with model-based recursive partitioning (MOB; Zeileis, Hothorn, & Hornik, 2008) in a unified framework so as to capture covariate-dependent heterogeneity efficiently. Classical recursive partitioning methods, such as classification and regression trees (Breiman, Friedman, Stone, & Olshen, 1984; Loh, 2011), focus on identifying subgroups involving different values of a response variable only. On the

other hand, MOB aims to fit a specified statistical model to each of heterogeneous subgroups identified successively based on an additional set of covariates. In this way, it can detect covariate-dependent subgroups that lead to different parameter estimates of the fitted statistical model (Seibold, Zeileis, & Hothorn, 2016; Strobl et al., 2015; Strobl, Wickelmaier, & Zeileis, 2011).

The proposed method, called MOB-ERA hereinafter, begins by fitting an ERA model to all observations, producing a single set of the ERA parameter estimates, and then successively inspects whether there are substantial changes in the estimated parameter values across covariate-dependent subgroups. This is achieved through the so-called parameter instability test in MOB that uses the individual contributions to the score function, as will be discussed in detail in the Methods section. The method provides a tree diagram that displays hierarchically a nested structure of all the covariates selected for partitioning. Each end node of the tree represents a non-overlapping subgroup that entails its own ERA parameter estimates. This tree can greatly aid in visualizing how the partitioning covariates interact with each other in a hierarchical manner and how each group can be characterized by combinations of these covariates.

The paper is organized as follows. We begin with an overview of ERA and present the proposed method, focusing on how MOB can be combined with ERA for finding covariate-dependent subgroups. We then conduct a simulation study to evaluate the performance of MOB-ERA. We apply the method to data from the 2012 National Survey on Drug Use and Health (NSDUH) concerning nicotine dependence among US adults and their associated predictors. This application shows the use of continuous and categorical sociodemographic covariates for subgroup identification in ERA. We conclude by briefly discussing the implications of the method and potential topics for future research.

Methods

Parametric ERA

Assume that there are K different sets of predictors, each of which consists of P_k predictors (k = 1, ⋯, K). Let x_{ikp} denote the ith value of the pth variable in the kth predictor set (i = 1, ⋯, N; p = 1, ⋯, P_k) and $\boldsymbol{x}_i = (x_{i11}, \cdots, x_{ikp})$ denote a 1 by P vector of predictors for the ith observation, where $P = \sum_{k=1}^{K} P_k$. Let y_i denote the ith value of the response variable. We assume that y_i follows an exponential family distribution with a mean μ_i and variance $\phi\sigma_i^2$, where ϕ is a constant dispersion parameter. Let w_{kp} denote a component weight assigned to x_{ikp} and $\boldsymbol{w}_k = (w_{k1}, \cdots, w_{kP_k})'$ denote a P_k by 1 vector of component weights in the kth predictor set. Let $f_{ik} = \sum_{p=1}^{P_k} x_{ikp} w_{kp}$ denote the ith component score of the kth component, which is the sum of weighted predictors for the ith observation in the kth predictor set. Let b_k denote the regression coefficient relating the kth component to the response variable. Let η_i and $g(\cdot)$ denote the ith linear predictor of y_i and a known link function that describes how μ_i is related to η_i, respectively. We assume that all the predictors are standardized with zero means and unit variances (Takane & Hwang, 2005).

The ERA model (Hwang et al., 2015; Lee et al., 2016) is then expressed as

$$\begin{aligned} g(\mu_i) &= \eta_i \\ &= \sum_{k=1}^{K} \left[\sum_{p=1}^{P_k} x_{ikp} w_{kp} \right] b_k = \boldsymbol{x}_i \mathbf{W} \boldsymbol{b} \\ &= \sum_{k=1}^{K} f_{ik} b_k = \boldsymbol{f}_i \boldsymbol{b}, \end{aligned} \tag{1}$$

where $\mathbf{W} = \begin{bmatrix} \mathbf{w}_1 & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & \mathbf{w}_K \end{bmatrix}$, $\boldsymbol{f}_i = (f_{i1}, \cdots, f_{iK})$, and $\boldsymbol{b} = (b_1, \cdots, b_K)'$. As shown in (1), each set of predictors reduces to a single component, which in turn influences the response variable. Each component weight w_{kp} shows the contribution of each predictor variable to obtaining its component as in canonical correlation analysis, whereas the regression coefficient b_k signifies the effect of each component on the response variable as in linear regression. In this regard, ERA carries out data reduction and linear regression simultaneously, as discussed earlier. Figure 1 displays an example of the ERA model, where a response variable is influenced by three components ($K = 3$), each of which is associated with two predictors ($P_1 = P_2 = P_3 = 2$). For this example, $\mathbf{W}$ and $\boldsymbol{b}$ are given as

$$\mathbf{W} = \begin{pmatrix} w_{11} & 0 & 0 \\ w_{12} & 0 & 0 \\ 0 & w_{21} & 0 \\ 0 & w_{22} & 0 \\ 0 & 0 & w_{31} \\ 0 & 0 & w_{32} \end{pmatrix} \text{ and } \boldsymbol{b} = \begin{pmatrix} b_1 \\ b_2 \\ b_3 \end{pmatrix}.$$

Insert Figure 1 around here

We assume a canonical link function $g(\cdot)$ that sets $\eta_i = \boldsymbol{x}_i \mathbf{W} \boldsymbol{b} = \boldsymbol{f}_i \boldsymbol{b}$. Then, the log-likelihood function of the ERA model for N observations can be written as

$$\ell(\boldsymbol{\theta}_{\text{ERA}}; y_1, \ldots, y_N) = \sum_{i=1}^{N} y_i \boldsymbol{x}_i \mathbf{W} \boldsymbol{b} - \beta(\boldsymbol{x}_i \mathbf{W} \boldsymbol{b}) = \sum_{i=1}^{N} y_i \boldsymbol{f}_i \boldsymbol{b} - \beta(\mathbf{f}_i \boldsymbol{b}), \quad (2)$$

where $\boldsymbol{\theta}_{\text{ERA}}$ denotes a (P+K) by 1 vector that stacks $\boldsymbol{w}_k$ and $\boldsymbol{b}$. The maximum-likelihood (ML) parameter estimates of a log-likelihood are typically obtained by iteratively reweighted least squares (IRLS) based on the Newton-Raphson optimization algorithm (McCullagh & Nelder, 1989, Chapter 2.5; Nelder & Wedderburn, 1972). For ERA, maximizing (2) via IRLS is equivalent to minimizing the following generalized least-squares criterion (Hwang et al., 2015; Lee et al., 2016)

$$\varphi_{(w_{kp},b_k)} = \sum_{i=1}^{N} \omega_i (z_i - \sum_{k=1}^{K} \Big[\sum\nolimits_{p=1}^{P_k} x_{ikp} w_{kp} \Big] b_k)^2 = \sum_{i=1}^{N} \omega_i (z_i - \sum_{k=1}^{K} f_{ik} b_k)^2 , \quad (3)$$

with respect to w_{kp} and b_k, subject to $\sum_{i=1}^{N} f_{ik}^2 = N$, where $\omega_i = (\partial\mu_i/\partial\eta_i)^2/\tau_i$, τ_i is the variance function value evaluated at μ_i, and z_i is the so-called adjusted response variable with elements $z_i = \eta_i + (y_i - \mu_i)/\omega_i$ (McCullagh & Nelder, 1989, Chapter 2). An iterative algorithm similar to the alternating least-squares algorithm was proposed to minimize (3) (Hwang et al., 2015; Lee et al., 2016). This algorithm yields the ML estimates of the ERA parameters and their asymptotic standard errors. Refer to the Appendix for a detailed description of the algorithm.

Recursive Partitioning of ERA

As discussed earlier, ERA currently has no standard method for capturing covariate-dependent heterogeneity in the model parameters, thus potentially ignoring subgroup-specific relationships between predictor and response variables. To identify heterogeneous subgroups based on a given set of covariates in ERA, we propose MOB-ERA that combines ERA with the general steps of the MOB algorithm. More specifically, the so-called parameter instability test in MOB (Seibold et al., 2016; Zeileis et al., 2008) is used to split the data recursively into disjoint subgroups (also called nodes) B_s (s = 1, ⋯, S), each of which contains its own ERA parameters including

component weights and regression coefficients. This test focuses on whether there are statistically significant changes in parameter estimates (i.e., parameter instabilities) across subgroups derived from a partitioning covariate, under the null hypothesis of parameter homogeneity. To be clear, the term 'covariate' refers to a variable that affects the direction and/or strength of the relation between predictor and response variables, which has been interchangeably used with the term 'moderator' (Arah, 2008; Bollen & Bauldry, 2011; Seibold et al., 2016; Thomas, Bornkamp, & Seibold, 2018).

Let $s\left(\boldsymbol{\theta}_{\text{ERA}}\right)$ denote the score function, i.e., the first derivative of the log-likelihood function in (2). Let $\hat{\psi}_i$ be the empirical contribution of the *i*th individual to the score function,

$$\hat{\psi}_i = s\left(\hat{\boldsymbol{\theta}}_{\text{ERA}}; y_i\right) = \left.\frac{\partial \ell(\boldsymbol{\theta}_{\text{ERA}}; y_i)}{\partial \boldsymbol{\theta}_{\text{ERA}}}\right|_{\hat{\boldsymbol{\theta}}_{\text{ERA}}}, \tag{4}$$

where $\hat{\boldsymbol{\theta}}_{\text{ERA}}$ denotes the ML parameter estimates at convergence. If only one set of parameters $\boldsymbol{\theta}_{\text{ERA}}$ holds for all *N* observations (i.e., no presence of covariate-dependent heterogeneity), then the empirical score contributions ($\hat{\psi}_i$; $i = 1, \cdots, N$) would fluctuate randomly around their mean (i.e., zero), regardless of how the observations are divided or grouped by a covariate. For example, let us consider "age" a partitioning covariate. After obtaining a set of the ERA parameter estimates over all *N* observations, we can sort their empirical score contributions, $\hat{\psi}_i$, by age. If no age-dependent heterogeneity is present, the ordered score contributions will not show any structural fluctuations over the entire range of age. But, in the presence of age-dependent heterogeneity, a systematic deviation of the ordered contributions from zero over the range of age will be observed.

This way of investigating the individual score contributions over the range of a covariate gives rise to several test statistics for the parameter instability test (see Merkle, Fan, & Zeileis, 2014; Zeileis & Hornik, 2007; Zeileis et al., 2008). All these statistics are based on the cumulative sum of the sorted empirical score contributions, the so-called empirical fluctuation process, and the exact form of the test statistic depends on whether the covariate is continuous (e.g., age), ordinal (e.g., education levels), or nominal (e.g., gender). For example, a test statistic for a continuous covariate is given by the maximum of the squared L_2 norm of the empirical fluctuation process scaled by its variance. Details of the parameter instability tests are discussed in Zeileis and Hornik (2007). The parameter instability test is performed for each and every covariate considered, and the observations are divided into subgroups if at least one of the partitioning covariates yields a p-value below the pre-specified significance level of α. The covariate associated with the smallest p-value is used as the partitioning variable at the current stage of data partitioning.

After choosing a covariate most associated with parameter instability, MOB-ERA determines a certain cutoff value (or a cut-point) in the selected covariate that makes two resulting subgroups of observations, say B_1 and B_2, as different as possible with respect to the estimated ERA parameters $\hat{\boldsymbol{\theta}}_{\text{ERA}}$. More specifically, for every conceivable value of the covariate, the sum of each subgroup's log-likelihood is calculated based on the ERA parameters estimated for the two groups, i.e., $\ell(\hat{\boldsymbol{\theta}}_{\text{ERA}}^{(B_1)}) + \ell(\hat{\boldsymbol{\theta}}_{\text{ERA}}^{(B_2)})$. The covariate value that maximizes the sum of the partitioned log-likelihoods is selected as the cut-point, leading to two subgroups of observations. Subsequently, within each of the subgroups, the same procedures of parameter instability test and cut-point selection are repeated until some stopping criteria met, as discussed in the next subsection.

Figure 2 displays an illustrative example of a MOB-ERA tree, where three subgroups (B_1, B_2, and B_3) of different sizes (n_1, n_2, and n_3) are identified based on two partitioning covariates (age and gender). Based on the ERA model in Figure 1, all observations are first partitioned into males and females. The male group (subgroup 3) involves no significant parameter instability by age, whereas the female group is further split by age, resulting in two more subgroups of women aged up to 30 (subgroup 1) and over 30 (subgroup 2). Each identified subgroup will provide its own ERA parameter estimates that are generally displayed in the boxes.

Insert Figure 2 around here

Pruning Strategy

In a recursive partitioning method, pruning is generally used to remove nodes to avoid overfitting (Strobl, Malley, & Tutz, 2009). In MOB-ERA, the following pre-pruning strategies are available: the data partitioning procedures are repeated until (a) no more covariate leads to statistically significant parameter instabilities, (b) a pre-specified threshold for the minimum number of observations left in a node is reached, or (c) all nodes are pure with respect to covariate values, where a pure node represents a subgroup that has observations belonging to the same covariate group. For large samples, however, these pre-pruning strategies may be less ideal because even a small degree of parameter instability can turn out to be statistically significant (Seibold et al., 2016; Zeileis et al., 2008).

MOB-ERA can also adopt the post-pruning strategy using information criteria, such as AIC or BIC, where pruning is started from the bottom of the tree upwards, removing one sub-

node at a time. For example, we may compare the following two AIC values to decide whether to prune a node:

$$\text{AIC}^{(\text{Parent node})} = -2 \cdot \ell(\hat{\boldsymbol{\theta}}_{\text{ERA}}^{(\text{Parent node})}) + 2 \cdot h^{(\text{Parent node})} \tag{5}$$

and

$$\text{AIC}^{(\text{Subsequent nodes: A and B})} = -2 \cdot (\ell(\hat{\boldsymbol{\theta}}_{\text{ERA}}^{(\text{Node A})}) + \ell(\hat{\boldsymbol{\theta}}_{\text{ERA}}^{(\text{Node B})})) + 2 \cdot (h^{(\text{Node A})} + h^{(\text{Node B})}), \tag{6}$$

where $\ell(\hat{\boldsymbol{\theta}}_{\text{ERA}}^{(\cdot)})$ denotes the log-likelihood of the ERA model evaluated at the estimated parameters, and h denotes the number of free parameters. The AIC in (5) represents the relative amount of information assuming a single set of parameter estimates (*simpler model of homogeneity*), where the AIC in (6) quantifies the information assuming different sets of parameter estimates (*complex model of heterogeneity*). For example, in Figure 2, assume that $\text{AIC}^{(\text{Node 2})} > \text{AIC}^{(B_1 \text{ and } B_2)}$. Then, the split of Node 2 into the subgroups B_1 and B_2 is kept in the final tree because this results in a smaller AIC value than the tree without these subgroups. By means of the pre- and/or post-pruning strategies, MOB-ERA can generate a hierarchy of selected covariates, which leads to heterogeneous subgroups of observations, in an automatic manner.

Simulation Study

We investigated a Type I error rate, power, and classification accuracy of MOB-ERA. In the MOB framework, a Type I error can be defined as the probability of having at least one split when none of the covariates are associated with parameter instabilities (Fokkema, Smits, Zeileis, Hothorn, & Kelderman, 2018; Seibold, Hothorn, & Zeileis, 2018; Wickelmaier & Zeileis, 2018). The Type I error performance of a new MOB extension has important practical implications because it is closely related to overfitting, where the tree partitions observations according to the

noise rather than the true covariate-dependent structure. Thus, we examined whether the Type I error rate was controlled across different simulation conditions. We also investigated how well and accurately MOB-ERA could detect parameter instability, thereby identifying the subgroups derived from pre-specified partitioning covariates correctly.

Simulation Design and Data Generation

We specified an ERA model that was composed of two components ($K = 2$) and a response variable. No correlation between the components was assumed. We fixed one regression coefficient b_1 to .3 but allowed the other regression coefficient b_2 to vary depending on how much of the variance in the response variable was accounted for by the two components (R^2). We considered three levels for the variance explained (R^2 = .2, .4, and .6), which in turn resulted in three different values of b_2 (b_2 = .33, .56, and .71). Each component was linked to four predictor variables ($P_k = 4$) with the pre-determined weight values, $\boldsymbol{w}_1 = (.7, .6, .5, .4)'$ and $\boldsymbol{w}_2 = (.6, .5, .4, .3)'$. The number of components and predictors remained the same over the different simulation conditions.

We considered six different sample sizes: $N = 90$, 120, 180, 300, 600, and 900. This total sample size N was then divided into three subgroups, whose sizes were denoted by n_1, n_2, and n_3, with respect to two partitioning covariates, Z_1 and Z_2. Z_1 was randomly sampled from a binomial distribution, $Z_1 \sim B(N, p_1)$, whereas Z_2 was from a uniform distribution between -1 and 1. Accordingly, three covariate-dependent subgroups were defined as follows:

$$\text{Total } N = \begin{cases} \text{Group1 } (n_1)\text{: if } Z_1 = 0 \\ \text{Group2 } (n_2)\text{: if } (Z_1 = 1) \wedge (Z_2 \leq 0) \\ \text{Group3 } (n_3)\text{: if } (Z_1 = 1) \wedge (Z_2 > 0). \end{cases}$$

The value of $p_1 = P(Z_1=0)$ was either 1/3 or 2/3 to generate two different subgroup size conditions—unbalanced and balanced subgroup sizes. Note that when $p_1 = 1/3$, the number of observations for each subgroup was, on average, $(n_1, n_2, n_3) = (1/3, 2/3\times1/2, 2/3\times1/2)N$, leading the number of observations to be all equal across the subgroups (balanced condition). When $p_1 = 2/3$, $(n_1, n_2, n_3) = (2/3, 1/3\times1/2, 1/3\times1/2)N$, resulting in the unbalanced condition where one group size was larger than the others. We also included a noise covariate Z_3 that was completely unrelated to the subgroups to examine whether MOB-ERA could accurately select the correct covariate when partitioning data. The noise covariate Z_3 was sampled from a uniform distribution between -1 and 1.

In this study, the degree of parameter instability in component weights was controlled by an instability control parameter $\delta = \{0, 0.1, 0.2, \text{or } 0.3\}$, the amount of deviation from the pre-determined weight values $\boldsymbol{w}_1$ and $\boldsymbol{w}_2$: $\mathbf{W}_{\text{Group1}} = \begin{bmatrix} \mathbf{w}_1 - \delta & \mathbf{0} \\ \mathbf{0} & \mathbf{w}_2 - \delta \end{bmatrix}$, $\mathbf{W}_{\text{Group2}} = \begin{bmatrix} \mathbf{w}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{w}_2 \end{bmatrix}$, and $\mathbf{W}_{\text{Group3}} = \begin{bmatrix} \mathbf{w}_1 + \delta & \mathbf{0} \\ \mathbf{0} & \mathbf{w}_2 + \delta \end{bmatrix}$. We used δ to generate either the homogeneity condition for evaluating the Type I error ($\delta = 0$) or the heterogeneity condition for evaluating power and accuracy ($\delta \neq 0$). Under the heterogeneity condition, we further differed the two regression coefficients by changing the signs (directions) as follows:

$$\boldsymbol{b} = \begin{cases} \boldsymbol{b}_{\text{Group1}} = (-b_1, +b_2)' \text{ if } Z_1=0 \\ \boldsymbol{b}_{\text{Group2}} = (+b_1, -b_2)' \text{ if } (Z_1=1) \wedge (Z_2 \leq 0) \\ \boldsymbol{b}_{\text{Group3}} = (+b_1, +b_2)' \text{ if } (Z_1=1) \wedge (Z_2 > 0). \end{cases}$$

Following the data generation approach of Becker, Rai, and Rigdon (2013), the variance-covariance matrix of the predictor and response variables, $\boldsymbol{\Sigma}$, was obtained based on the ERA parameters described above. We generated 1000 datasets from a multivariate normal distribution

with zero means and $\boldsymbol{\Sigma}$ for each combination of variance explained (R^2), sample size (N), parameter homogeneity or heterogeneity (δ), and number of observations across subgroups (balanced or unbalanced). We applied MOB-ERA to the datasets to compute its empirical Type I error rate, power, and classification accuracy under each condition. All data generation and computations were carried out using the R system for statistical computing version 3.5.1. We wrote an R code to implement ERA, which is archived on GitHub at https://github.com/generalizedERA. We used the "lmtree" function of the R package "partykit" (version 1.2-5; Hothorn & Zeileis, 2015) for the parameter instability test and cup-point selection.

Results

In this study, an empirical Type I error was calculated by counting how many of the samples were falsely partitioned under the homogeneity condition ($\delta = 0$). Table 1 presents the empirical Type I error rates across the different sample sizes and the different values of R^2. In all the conditions, MOB-ERA tended to produce somewhat conservative Type I error rates, i.e., yielded smaller values than the nominal significance level of .05, and this pattern became more apparent in smaller samples ($N < 300$). This is consistent with previous MOB studies (e.g., Frick, Strobl, & Zeileis, 2014; Seibold et al., 2018), in which the parameter instability test in MOB with many partitioning covariates were often shown to be conservative, especially in small samples, because of the Bonferroni correction applied. In large samples ($N \geq 300$), however, MOB-ERA seemed to control Type I errors reasonably well regardless of the value of R^2 and remain close to the nominal significance level of .05. To hold the nominal Type I error rate, therefore, it may be important to ensure a sufficiently large number of observations relative to the number of

partitioning covariates considered for the parameter instability test, e.g., at least 300 observations for three partitioning covariates in this study.

Insert Table 1 around here

Table 1 also provides the empirical power of MOB-ERA for different combinations of sample sizes, δ, and R^2 values under the heterogeneity condition, i.e., when the null hypothesis of parameter stability was not true. For the calculation of the empirical power, we counted how many times the parameter instability test was turned out to be significant, so that a sample was correctly partitioned by Z_1 and/or Z_2 out of 1000 random samples. As shown in the table, the empirical power estimates tended to increase when the sample size and/or R^2 increased. More specifically, the influence of the sample size or R^2 on the power was strongly dependent on the number of observations for each subgroup: Under the balanced condition, MOB-ERA was able to detect instabilities beyond a power threshold of .9 across all the sample sizes and R^2 values. Under the unbalanced condition, conversely, the power dropped quickly in small samples ($N \leq$ 120) even when the difference in the magnitude of regression coefficients between groups was large (e.g., $R^2 = .6$). The influence of the component weight instability (i.e., the magnitude of δ) on the power was minimal until the δ value approached 0.3. Thus, to ensure an adequate level of power of MOB-ERA in small samples, the size of any subgroup should not be too dominant. Although not reported in Table 1, we found that the estimated probability that a sample was erroneously partitioned by the noise covariate Z_3 was zero across all the conditions.

Finally, the classification accuracy of subgroup memberships was measured using the Cramér's V, which is a normalized χ^2 statistics of true and predicted group memberships in a

cross-table (Mirkin, 2001). It ranges between 0 and 1, where 1 means complete match between true and predicted subgroup memberships. Table 1 also displays the average Cramér's V values for the different sample sizes, δ, and R^2 values under the heterogeneity condition. Under the balanced condition, on average, the Cramér's V increased with the sample size, δ, and R^2. Moreover, Cramér's V were all over .9 even in small samples, which indicates a high level of accuracy in recovering the true subgroup memberships. Under the unbalanced conditions, Cramér's V decreased when the sample size and R^2 were small. This is expected because the row totals in a cross-table are extremely uneven when one group size is much larger than the others, leading to exaggerated V estimates (Mirkin, 2001). Conversely, the V estimates almost approached 1 when the sample size increased ($N > 180$) and the values of δ and R^2 became large. Interestingly, Cramér's V decreased again when $N = 900$ because MOB-ERA ended up further partitioning the largest subgroup into more than the pre-specified one. This suggests that post-pruning might be necessary in large samples to avoid such overfitting, especially when a group is dominant in size.

Empirical Application

We applied MOB-ERA to public data collected from the 2012 National Survey on Drug Use and Health (NSDUH) (United States Department of Health and Human Services, Substance Abuse and Mental Health Services Administration [SAMHSA], 2013). This survey was conducted from January through December 2012 and interviewed a number of residents aged 12 and older in American households. The respondents were asked to answer various questions concerning their use of substances (e.g., tobacco, alcohol, marijuana, etc.), mental and physical health issues, and sociodemographic characteristics (e.g., age, gender, ethnicity, marital status, etc.).

In this application, we attempted to examine sociodemographic differences in the effects of predictors related to early exposure to substances, mental health, and SES on nicotine dependence among US adults. The response variable, the degree of nicotine dependence, was the average score of the Nicotine Dependence Syndrome Scale (SAMHSA, 2013). We identified a total of 11 predictors that were available in the 2012 NSDUH data based on previous studies concerning the predictors of nicotine dependence on samples of US adults (e.g., Bohadana, Nilsson, Martinet, & Rasmussen, 2003; Breslau, Fenn, & Peterson, 1993; Breslau, Kilbey, & Andreski, 1994; Daeppen et al., 2000; Green, Jucha, & Luz, 1986; Hu et al., 2006; Jackson, Knight, & Rafferty, 2010; Kandel, Chen, Warner, Kessler, & Grant, 1997; Kandel & Chen, 2000; Khuder, Dayal, & Mutgi, 1999; Schmitz, Kruse, & Kugler, 2003). Then, the predictors were grouped into three categories, such as substance initiation age (F_1), mental health status (F_2), and SES (F_3), which were represented as components in the ERA model. Table 2 presents a description of all the variables and their summary statistics. It also shows which component is associated with which predictors. Figure 3 displays the specified ERA model, where three sets of predictors related to F_1, F_2, and F_3 were to influence the degree of nicotine dependence. The number of respondents was $N = 8{,}412$ in our analysis.

Insert Table 2 and Figure 3 around here

The use of an independent hold-out dataset (often called a test or validation set) for model evaluation has been emphasized in many contexts, especially in the recursive partitioning literature (Bauer & Kohavi, 1999; Elith, Leathwick, & Hastie, 2008; Hastie, Tibshirani, & Friedman, 2009). Thus, we divided the dataset randomly into two disjoint sub-datasets—training

(N_{train} = 4,206) and test (N_{test} = 4,206) datasets. We used the test set to validate the generalizability of our MOB-ERA results obtained from the training set.

As partitioning covariates, we considered four sociodemographic variables: age, gender, marital status, and ethnicity. Refer to Table 2 for their summary statistics. Many previous studies have reported several subgroups of nicotine dependence that could be differentiated by age, gender, or ethnicity (e.g., Bohadana et al., 2003; Breslau et al., 1993; Daeppen et al., 2000; Hu et al., 2006; Jackson et al., 2010; Kandel et al., 1997; Kandel & Chen, 2000; Khuder et al., 1999). In these studies, covariate-dependent subgroups were pre-defined by researchers (e.g., females vs. males, Black vs. White smokers, etc.). However, in practice, it is often unclear how and which covariates may interact with each other, and difficult to determine such subgroups in advance, especially when there are continuous covariates, categorical covariates with multiple levels, and/or a number of covariates at the same time (Strobl et al., 2015; Su et al., 2009; Zeileis et al., 2008).

As stated earlier, the final MOB-ERA model can be decided by pre- and post-pruning to avoid potential overfitting. The following pruning procedures were the same for both training and test sets: When splitting the data, the tree size was determined by the parameter instability tests (i.e., data splitting is continued until no covariate was statistically significant at α = .05) and the minimal node size of 500 (pre-pruning). Considering the large number of respondents, we then pruned the tree afterwards using the AIC-based pruning function (post-pruning). Figure 4 presents the final MOB-ERA solutions obtained from the training and test sets. In the figure, the internal nodes, represented by circles, show which and how covariates partition the data into subgroups in a hierarchical manner. Each circle shows the selected covariate and its *p*-value obtained from the parameter instability test, as will be further discussed shortly. Each grey box at

the bottom denotes a leave or terminal node of the tree, representing a subgroup identified. It also displays the number of respondents and the estimated regression coefficients of each subgroup. Node number is given at the top of every circle and box.

Insert Figure 4 around here

Table 3 summarizes the results of the parameter instability tests. Each node in the table shows the values of the test statistics and *p*-values for each of the four covariates. A node was partitioned into subgroups when at least one covariate was statistically significant at $\alpha = .05$ (until the minimum node size of 500 was reached). The covariate with the smallest *p*-value is used as the partitioning variable at each node. In the training set, ethnicity was selected as the first partitioning covariate (Node 1), splitting them into two groups—Whites and all the other ethnicities (Hispanic and Non-Hispanic-All). For the group of Whites, two age groups (i.e., up to 22.5 and over 22.5) were found to be significantly different (Node 3), whereas for all other ethnicities, no further split was carried out. As shown in the table (and also displayed in Figure 4), the final hierarchy of the partitioning covariates was the same for both training and test sets. This suggests that, using the pre- and post-pruning strategies, MOB-ERA could reliably identify heterogeneous subgroups of nicotine dependence based on the partitioning covariates.

Insert Table 3 around here

Table 4 presents the estimated component weights, their standard errors, and *p*-values for the identified subgroups in Nodes 2, 4, and 5. The first three columns of the table show the results obtained from the training set. As shown in the table, the component weight estimate for

age of first cigarette use (w_{11}) was positive and statistically significant across all three subgroups, indicating that cigarette initiation contributed to forming F_1, substance initiation age, in explaining the degree of nicotine dependence. Alcohol and marijuana initiation age were also statistically significant in the group of young Whites aged up to 22.5 (Node 4). With regard to the second predictor set related to F_2, mental health status, different predictors were statistically significant among the subgroups: the functional impairment level in daily life (w_{22}) in the non-White respondents (Node 2), the overall level of nonspecific psychological distress (w_{21}) in young Whites (Node 4), and the history of serious suicidal ideation (w_{23}) in older Whites aged over 22.5 (Node 5). In the last predictor set, the weight estimate for education level (w_{31}) was statistically significant, contributing to determining F_3, SES, across all the subgroups. In addition, the health insurance status (w_{32}) was also significant in Nodes 2 and 4, whereas the job status (w_{34}) was significant in Node 5. As shown in the table, many of the estimated component weights turned out to be statistically insignificant. This is, however, common in practice when dealing with a large number of predictors simultaneously (DeSarbo, Hwang, Blank, & Kappe, 2015; Kim et al., 2020; Lee et al., 2016). Lastly, as provided in the last three columns of the table, similar results were obtained from the test set.

Insert Table 4 around here

Table 5 shows the estimated regression coefficients and their standard errors per subgroup. The estimates are also displayed at each terminal node in Figure 4. Note that we can compare the relative magnitudes of the regression coefficient estimates because they are standardized ones in ERA. As shown in the table, earlier substance use (F_1), worse mental health

(F_2), and lower SES (F_3) were associated with a higher level of nicotine dependence in all identified subgroups. However, the magnitudes of their effects varied across the groups. For example, earlier substance use had a larger effect on nicotine dependence in the group of young Whites aged up to 22.5 (Node 4), compared to the other groups. SES had the smallest effect on the nicotine dependence in the non-White respondents (Node 2), whereas it had larger effects in White respondents (Nodes 4 and 5). Moreover, the group of older Whites showed the largest effect of mental health status on nicotine dependence among the three groups. Again, similar findings were obtained from the test set.

Insert Table 5 around here

Concluding Remarks

We combined ERA with MOB to identify potentially heterogeneous subgroups of observations based on a set of auxiliary covariates in the context of ERA. The proposed method, MOB-ERA, successively repeats the procedures of probing parameter instabilities and finding a cut-point for covariates, given a specified ERA model. This results in a tree diagram that displays covariate-dependent characteristics of identified subgroups, facilitating an understanding of subgroup-specific effects of components on a response variable. Unlike conventional multiple-group analyses, where grouping covariates are pre-specified ahead of data analysis, MOB-ERA detects a meaningful combination of covariates upon the completion of the recursive partitioning procedure.

The simulation study showed that MOB-ERA seemed to control for the Type I error rate reasonably well over the whole range of the simulation conditions considered. The relatively conservative level of Type I error rates in small samples became close to the nominal level of .05 when the sample size became large. The empirical power and classification accuracy also showed that MOB-ERA satisfactorily recovered the predefined heterogeneous subgroups particularly when the number of observations was roughly equal across the subgroups.

We also demonstrated how MOB-ERA could identify covariate-dependent heterogeneous subgroups, using a well-known national survey dataset in the US. When partitioning the data based on the specified ERA model, we applied both pre- and post-pruning strategies to avoid overfitting and enhance the generalizability of the resulting MOB-ERA tree. The final hierarchy of partitioning covariates was derived in a data-driven manner, without needing to specify in advance which covariates should be included and how they interact with each other. The combination of the selected covariates in the final MOB-ERA tree resulted in socio-demographically diverse subgroups, each of which showed different strengths of component effects on the response variable. Moreover, the findings obtained from a random half of the dataset (a training set) were much the same as those from the other half (an independent validation set), suggesting that MOB-ERA was reliable in detecting heterogeneous subgroups.

As with many other recursive partitioning methods, a major limitation of MOB-ERA is that its single-tree solution can be highly variable, i.e., the hierarchy of partitioning structure can be changed entirely by a small change in training data (Garge, Bobashev, & Eggleston, 2013; Strobl et al., 2009). It may be necessary to technically refine the method to alleviate this potential variability problem of a single MOB-ERA tree. For example, we may combine the proposed method into the frameworks of bagging (Breiman, 1996) or random forests (Breiman, 2001).

These so-called ensemble methods build a large number of separate trees and average them to improve generalizability of a single tree estimator. Both bagging and random forests fit trees independently to random samples of the original training dataset, where the random sampling procedure is carried out either using bootstrapping (i.e., sampling with replacement of the same size) or subsampling (i.e., sampling without replacement of smaller size). Random forests also include random selection of predictors to prevent some predominant predictors from being repeatedly selected across random trees. Adopting these ensemble methods to MOB-ERA may help enhance the generalizability and predictive performance of a single MOB-ERA tree, which warrants future research.

Lastly, it would be worthwhile to further investigate the generalizability of the final solution of MOB-ERA. In many social and behavioral studies, it is often assumed that a sample at hand (i.e., training data) is a good reflection of what will be encountered in future data; thus, the final model is selected as the one optimized on the training data. But when researchers aim to develop a model that can better assist decision-making in future unseen data, it would be crucial to assess a model's performance on an "out-of-sample" (i.e., independent data not used for model development). In the Empirical Application section, we randomly split the data into two separate subsets and obtained very similar findings on both sets. However, this split-sample approach can be limited in two ways. First, it is inefficient when sample size is small, leading to training and validation datasets, both of which are small. Second, high variability in the final model selection can be introduced because of its reliance on a single split of the data. Thus, in future research, we may consider applying resampling methods, such as cross validation or the bootstrap method, to evaluate the performance of MOB-ERA models as alternatives to the split-

sample approach. This may be of particular use for researchers who seek to build a model that generalizes to unseen samples, especially when overfitting is of concern.

References

Arah, O. A. (2008). The role of causal reasoning in understanding Simpson's paradox, Lord's paradox, and the suppression effect: covariate selection in the analysis of observational studies. *Emerging Themes in Epidemiology*, *5*(1), 5. https://doi.org/10.1186/1742-7622-5-5

Bauer, E., & Kohavi, R. (1999). An Empirical Comparison of Voting Classification Algorithms: Bagging, Boosting, and Variants. *Machine Learning*, *36*(1–2), 105–139. https://doi.org/10.1023/A:1007515423169

Becker, J.-M., Rai, A., & Rigdon, E. (2013). Predictive validity and formative measurement in structural equation modeling: Embracing practical relevance. In *the International Conference on Information Systems (ICIS)*. Retrieved from https://scholarworks.gsu.edu/marketing_facpub

Bohadana, A., Nilsson, F., Martinet, Y., & Rasmussen, T. (2003). Gender differences in quit rates following smoking cessation with combination nicotine therapy: Influence of baseline smoking behavior. *Nicotine & Tobacco Research*, *5*(1), 111–116. https://doi.org/10.1080/1462220021000060482

Bollen, K. A., & Bauldry, S. (2011). Three Cs in measurement models: Causal indicators, composite indicators, and covariates. *Psychological Methods*, *16*(3), 265–284. https://doi.org/10.1037/a0024448

Brandmaier, A. M., von Oertzen, T., McArdle, J. J., & Lindenberger, U. (2013). Structural equation model trees. *Psychological Methods*, *18*(1), 71–86. https://doi.org/10.1037/a0030001

Breiman, L. (1996). Bagging Predictors. *Machine Learning*, *24*(2), 123–140. https://doi.org/10.1023/A:1018054314350

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). *Classification and regression trees*. Chapman and Hall/CRC.

Breslau, N., Fenn, N., & Peterson, E. L. (1993). Early smoking initiation and nicotine dependence in a cohort of young adults. *Drug and Alcohol Dependence*, *33*(2), 129–137. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/8261877

Breslau, N., Kilbey, M. M., & Andreski, P. (1994). DSM-III-R nicotine dependence in young adults: prevalence, correlates and associated psychiatric disorders. *Addiction (Abingdon, England)*, *89*(6), 743–754. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/8069175

Cauffman, E., & MacIntosh, R. (2006). A Rasch Differential Item Functioning Analysis of the Massachusetts Youth Screening Instrument. *Educational and Psychological Measurement*, *66*(3), 502–521. https://doi.org/10.1177/0013164405282460

Choi, J. Y., Kyung, M., Hwang, H., & Park, J.-H. (2019). Bayesian Extended Redundancy Analysis: A Bayesian Approach to Component-based Regression with Dimension Reduction. *Multivariate Behavioral Research*, 1–19. https://doi.org/10.1080/00273171.2019.1598837

Daeppen, J. B., Smith, T. L., Danko, G. P., Gordon, L., Landi, N. A., Nurnberger, J. I., … Schuckit, M. A. (2000). Clinical correlates of cigarette smoking and nicotine dependence in alcohol-dependent men and women. The Collaborative Study Group on the Genetics of Alcoholism. *Alcohol and Alcoholism*, *35*(2), 171–175. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/10787393

Daza, P., Cofta-Woerpel, L., Mazas, C., Fouladi, R. T., Cinciripini, P. M., Gritz, E. R., & Wetter,

D. W. (2006). Racial and Ethnic Differences in Predictors of Smoking Cessation. *Substance Use & Misuse*, *41*(3), 317–339. https://doi.org/10.1080/10826080500410884

DeSarbo, W. S., Hwang, H., Blank, A., & Kappe, E. (2015). Constrained Stochastic Extended Redundancy Analysis. *Psychometrika*, *80*(2), 516–534. https://doi.org/10.1007/s11336-013-9385-6

Elith, J., Leathwick, J. R., & Hastie, T. (2008). A working guide to boosted regression trees. *Journal of Animal Ecology*, *77*(4), 802–813. https://doi.org/10.1111/j.1365-2656.2008.01390.x

Fleishman, J. A., Spector, W. D., & Altman, B. M. (2002). Impact of differential item functioning on age and gender differences in functional disability. *The Journals of Gerontology. Series B, Psychological Sciences and Social Sciences*, *57*(5), S275-84. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/12198107

Fokkema, M., Smits, N., Zeileis, A., Hothorn, T., & Kelderman, H. (2018). Detecting treatment-subgroup interactions in clustered data with generalized linear mixed-effects model trees. *Behavior Research Methods*, *50*(5), 2016–2034. https://doi.org/10.3758/s13428-017-0971-x

Frick, H., Strobl, C., & Zeileis, A. (2014). To split or to mix? Tree vs. mixture models for detecting subgroups. In M. Gilli, G. González-Rodríguez, & A. Nieto-Reyes (Eds.), *COMPSTAT 2014 – 21st international conference on computational statistics* (pp. 379–386). Geneva: The International Statistical Institute/International Association for Statistical Computing. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.670.5743&rep=rep1&type=pdf#page=397

Garge, N. R., Bobashev, G., & Eggleston, B. (2013). Random forest methodology for model-

based recursive partitioning: the mobForest package for R. *BMC Bioinformatics*, *14*(1), 125. https://doi.org/10.1186/1471-2105-14-125

Green, M. S., Jucha, E., & Luz, Y. (1986). Blood pressure in smokers and nonsmokers: epidemiologic findings. *American Heart Journal*, *111*(5), 932–940. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/3706114

Hastie, T., Tibshirani, R., & Friedman, J. H. (2009). *The elements of statistical learning : data mining, inference, and prediction* (2nd ed.). Springer.

Hothorn, T., & Zeileis, A. (2015). partykit: A Modular Toolkit for Recursive Partytioning in R. *Journal of Machine Learning Research*, *16*(118), 3905–3909. Retrieved from http://jmlr.org/papers/v16/hothorn15a.html

Hu, M.-C., Davies, M., & Kandel, D. B. (2006). Epidemiology and correlates of daily smoking and nicotine dependence among young adults in the United States. *American Journal of Public Health*, *96*(2), 299–308. https://doi.org/10.2105/AJPH.2004.057232

Hwang, H., Suk, H. W., Takane, Y., Lee, J., & Lim, J. (2015). Generalized functional extended redundancy analysis. *Psychometrika*, *80*(1), 101–125. https://doi.org/10.1007/S11336-013-9373-X

Jackson, J. S., Knight, K. M., & Rafferty, J. A. (2010). Race and unhealthy behaviors: chronic stress, the HPA axis, and physical and mental health disparities over the life course. *American Journal of Public Health*, *100*(5), 933–939. https://doi.org/10.2105/AJPH.2008.143446

Kandel, D. B., & Chen, K. (2000). Extent of smoking and nicotine dependence in the United States: 1991-1993. *Nicotine & Tobacco Research : Official Journal of the Society for Research on Nicotine and Tobacco*, *2*(3), 263–274. Retrieved from

http://www.ncbi.nlm.nih.gov/pubmed/11082827

Kandel, D. B., Chen, K., Warner, L. A., Kessler, R. C., & Grant, B. (1997). Prevalence and demographic correlates of symptoms of last year dependence on alcohol, nicotine, marijuana and cocaine in the U.S. population. *Drug and Alcohol Dependence*, *44*(1), 11–29. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/9031816

Kandel, D. B., Kiros, G.-E., Schaffran, C., & Hu, M.-C. (2004). Racial/ethnic differences in cigarette smoking initiation and progression to daily smoking: a multilevel analysis. *American Journal of Public Health*, *94*(1), 128–135. https://doi.org/10.2105/ajph.94.1.128

Khuder, S. A., Dayal, H. H., & Mutgi, A. B. (1999). Age at smoking onset and its effect on smoking cessation. *Addictive Behaviors*, *24*(5), 673–677. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/10574304

Kim, S., Lee, S., Cardwell, R. L., Kim, Y., Park, T., & Hwang, H. (2020). An application of regularized extended redundancy analysis via generalized estimating equations to the study of co-occurring substance use among US adults. In *Quantitative Psychology. IMPS 2019.*

Lee, S., Choi, S., Kim, Y. J., Kim, B.-J., T2d-Genes Consortium, Hwang, H., & Park, T. (2016). Pathway-based approach using hierarchical components of collapsed rare variants. *Bioinformatics*, *32*(17), i586–i594. https://doi.org/10.1093/bioinformatics/btw425

Lee, S., Kim, S., Kim, Y., Oh, B., Hwang, H., & Park, T. (2019). Pathway analysis of rare variants for the clustered phenotypes by using hierarchical structured components analysis. *BMC Medical Genomics*, *12*, 100. https://doi.org/10.1186/s12920-019-0517-4

Lee, S., Kim, Y., Choi, S., Hwang, H., & Park, T. (2018). Pathway-based approach using hierarchical components of rare variants to analyze multiple phenotypes. *BMC Bioinformatics*, *19*(S4), 79. https://doi.org/10.1186/s12859-018-2066-9

Loh, W.-Y. (2011). Classification and regression trees. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *1*(1), 14–23. https://doi.org/10.1002/widm.8

Maher, E. (2004). Health-related quality of life of severely obese children and adolescents. *Child: Care, Health and Development*, *30*(1), 94–95. https://doi.org/10.1111/j.1365-2214.2004.t01-10-00388.x

McArdle, J. J., & Epstein, D. (1987). Latent growth curves within developmental structural equation models. *Child Development*, *58*(1), 110–133. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/3816341

McCullagh, P., & Nelder, J. A. (1989). *Generalized linear models* (2nd ed.). Chapman and Hall.

Merkle, E. C., Fan, J., & Zeileis, A. (2014). Testing for Measurement Invariance with Respect to an Ordinal Variable. *Psychometrika*, *79*(4), 569–584. https://doi.org/10.1007/s11336-013-9376-7

Merkle, E. C., & Zeileis, A. (2013). Tests of Measurement Invariance Without Subgroups: A Generalization of Classical Methods. *Psychometrika*, *78*(1), 59–82. https://doi.org/10.1007/s11336-012-9302-4

Mirkin, B. (2001). Eleven Ways to Look at the Chi-Squared Coefficient for Contingency Tables. *The American Statistician*, *55*(2), 111–120. https://doi.org/10.1198/000313001750358428

Nelder, J. A., & Wedderburn, R. W. M. (1972). Generalized Linear Models. *Journal of the Royal Statistical Society. Series A (General)*, *135*(3), 370–384. https://doi.org/10.2307/2344614

Raudenbush, S. W. (1997). Statistical analysis and optimal design for cluster randomized trials. *Psychological Methods*, *2*(2), 173–185. https://doi.org/10.1037/1082-989X.2.2.173

Richards, F. S. (1961). A method of maximum-likelihood estimation. *Journal of the Royal Statistical Society: Series B (Methodological)*, *23*(2), 469–475. Retrieved from

https://www.jstor.org/stable/pdf/2984037.pdf

Robinson, L., Murray, D., Alfano, C., Zbikowski, S., Blitstein, J., & Klesges, R. (2006). Ethnic differences in predictors of adolescent smoking onset and escalation: A longitudinal study from 7th to 12th grade. *Nicotine & Tobacco Research*, *8*(2), 297–307. https://doi.org/10.1080/14622200500490250

Royston, P., & Sauerbrei, W. (2004). A new approach to modelling interactions between treatment and continuous covariates in clinical trials by using fractional polynomials. *Statistics in Medicine*, *23*(16), 2509–2525. https://doi.org/10.1002/sim.1815

Schmitz, N., Kruse, J., & Kugler, J. (2003). Disabilities, Quality of Life, and Mental Disorders Associated With Smoking and Nicotine Dependence. *American Journal of Psychiatry*, *160*(9), 1670–1676. https://doi.org/10.1176/appi.ajp.160.9.1670

Seibold, H., Hothorn, T., & Zeileis, A. (2018). Generalised linear model trees with global additive effects. *Advances in Data Analysis and Classification*, 1–23. https://doi.org/10.1007/s11634-018-0342-1

Seibold, H., Zeileis, A., & Hothorn, T. (2016). Model-Based Recursive Partitioning for Subgroup Analyses. *The International Journal of Biostatistics*, *12*(1), 45–63. https://doi.org/10.1515/ijb-2015-0032

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). A Critical Assessment of Our Assumption. In *Experimental and quasi-experimental designs for generalized causal inference*. Boston, MA, US: Houghton: Mifflin and Company. Retrieved from https://psycnet.apa.org/record/2002-17373-000

Smith, L. L., & Reise, S. P. (1998). Gender differences on negative affectivity: an IRT study of differential item functioning on the Multidimensional Personality Questionnaire Stress

Reaction Scale. *Journal of Personality and Social Psychology*, *75*(5), 1350–1362. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/9866192

Strobl, C., Kopf, J., & Zeileis, A. (2015). Rasch Trees: A New Method for Detecting Differential Item Functioning in the Rasch Model. *Psychometrika*, *80*(2), 289–316. https://doi.org/10.1007/s11336-013-9388-3

Strobl, C., Malley, J., & Tutz, G. (2009). An introduction to recursive partitioning: rationale, application, and characteristics of classification and regression trees, bagging, and random forests. *Psychological Methods*, *14*(4), 323–348. https://doi.org/10.1037/a0016973

Strobl, C., Wickelmaier, F., & Zeileis, A. (2011). Accounting for Individual Differences in Bradley-Terry Models by Means of Recursive Partitioning. *Journal of Educational and Behavioral Statistics*, *36*(2), 135–153. https://doi.org/10.3102/1076998609359791

Su, X., Tsai, C.-L., Wang, H., Nickerson, D. M., & Li, B. (2009). Subgroup analysis via recursive partitioning. *Journal of Machine Learning Research*, *10*, 141–158. Retrieved from http://www.jmlr.org/papers/volume10/su09a/su09a.pdf

Takane, Y., & Hwang, H. (2005). An extended redundancy analysis and its applications to two practical examples. *Computational Statistics & Data Analysis*, *49*, 785–808. https://doi.org/10.1016/j.csda.2004.06.004

Tan, T., Choi, J. Y., & Hwang, H. (2015). Fuzzy Clusterwise Functional Extended Redundancy Analysis. *Behaviormetrika*, *42*(1), 37–62. https://doi.org/10.2333/bhmk.42.37

Thomas, M., Bornkamp, B., & Seibold, H. (2018). Subgroup identification in dose-finding trials via model-based recursive partitioning. *Statistics in Medicine*, *37*(10), 1608–1624. https://doi.org/10.1002/sim.7594

Von Stumm, S., & Plomin, R. (2015). Socioeconomic status and the growth of intelligence from

infancy through adolescence. *Intelligence*, *48*, 30–36. https://doi.org/10.1016/J.INTELL.2014.10.002

Wake, M., Salmon, L., Waters, E., Wright, M., & Hesketh, K. (2002). Parent-reported health status of overweight and obese Australian primary school children: a cross-sectional population survey. *International Journal of Obesity*, *26*(5), 717–724. https://doi.org/10.1038/sj.ijo.0801974

Wickelmaier, F., & Zeileis, A. (2018). Using recursive partitioning to account for parameter heterogeneity in multinomial processing tree models. *Behavior Research Methods*, *50*(3), 1217–1233. https://doi.org/10.3758/s13428-017-0937-z

Williams, J., Wake, M., Hesketh, K., Maher, E., & Waters, E. (2005). Health-Related Quality of Life of Overweight and Obese Children. *JAMA*, *293*(1), 70–76. https://doi.org/10.1001/jama.293.1.70

Yee, T. W., & Hastie, T. J. (2003). Reduced-rank vector generalized linear models. *Statistical Modelling: An International Journal*, *3*(1), 15–41. https://doi.org/10.1191/1471082X03st045oa

Zeileis, A., & Hornik, K. (2007). Generalized M-fluctuation tests for parameter instability. *Statistica Neerlandica*, *61*(4), 488–508. https://doi.org/10.1111/j.1467-9574.2007.00371.x

Zeileis, A., Hothorn, T., & Hornik, K. (2008). Model-Based Recursive Partitioning. *Journal of Computational and Graphical Statistics*, *17*(2), 492–514. https://doi.org/10.1198/106186008X319331

Zeller, M. H., & Modi, A. C. (2006). Predictors of Health-Related Quality of Life in Obese Youth. *Obesity*, *14*(1), 122–130. https://doi.org/10.1038/oby.2006.15

Table 1. Type I error, power, and Cramer's V coefficients under different sample and subgroup sizes obtained from the proposed method.

				Total Sample Size (N)					
Measures	Subgroup Sizes	δ	R^2	90	120	180	300	600	900
Type I error	(N/A)	(N/A)	.2	.01	.03	.03	.04	.04	.04
			.4	.02	.02	.03	.04	.04	.04
			.6	.02	.03	.03	.04	.05	.05
Power	Balanced	0.1	.2	.91	.91	.93	1.00	1.00	1.00
			.4	.93	.94	1.00	1.00	1.00	1.00
			.6	.93	.98	1.00	1.00	1.00	1.00
		0.2	.2	.91	.91	.93	1.00	1.00	1.00
			.4	.93	.95	1.00	1.00	1.00	1.00
			.6	.93	.98	1.00	1.00	1.00	1.00
		0.3	.2	.94	.96	1.00	1.00	1.00	1.00
			.4	.96	1.00	1.00	1.00	1.00	1.00
			.6	.98	1.00	1.00	1.00	1.00	1.00
	Unbalanced	0.1	.2	.11	.12	.90	1.00	1.00	1.00
			.4	.10	.11	.96	1.00	1.00	1.00
			.6	.11	.12	1.00	1.00	1.00	1.00
		0.2	.2	.10	.11	.91	1.00	1.00	1.00
			.4	.11	.12	.96	1.00	1.00	1.00
			.6	.12	.12	1.00	1.00	1.00	1.00
		0.3	.2	.12	.12	.90	1.00	1.00	1.00
			.4	.15	.15	1.00	1.00	1.00	1.00
			.6	.21	.22	1.00	1.00	1.00	1.00
Cramer's V	Balanced	0.1	.2	.90	.90	.94	.97	.98	.99
			.4	.91	.91	.95	.98	.98	.99
			.6	.93	.94	.99	.99	.99	.99
		0.2	.2	.90	.91	.95	.97	.98	.99
			.4	.91	.91	.95	.98	.98	.99
			.6	.94	.94	.99	.99	.99	.99
		0.3	.2	.91	.91	.95	.98	.99	.99
			.4	.91	.93	.98	.99	.99	.99
			.6	.95	.95	.99	.99	.99	.99
	Unbalanced	0.1	.2	.79	.83	.84	.91	.93	.90
			.4	.82	.83	.83	.92	.95	.89
			.6	.83	.84	.84	.95	.96	.90
		0.2	.2	.81	.83	.84	.92	.95	.88
			.4	.81	.84	.85	.92	.96	.89
			.6	.83	.84	.85	.99	.99	.90
		0.3	.2	.85	.85	.85	.95	.96	.90
			.4	.86	.86	.86	.99	.99	.90
			.6	.86	.86	.90	.99	.99	.90

Table 2. A description of variables and summary statistics for the 2012 NSDUH data (N=8,412)

Variable Names	Measures (Range or Categories)	Mean (Q1, Q3) [a]
Response Variable		
Nicotine (cigarette) dependence	Average score over 17 items of the Nicotine Dependence Syndrome Scale (1-5)	2.55 (2, 3)
Predictors		
$\mathbf{F}_1$: Substance initiation age		
Cigarette (Cig)	Age of first cigarette use	15.81 (14, 18)
Alcohol (Alc)	Age of first alcohol use	16.82 (15, 18)
Marijuana (Mar)	Age of first marijuana use	16.94 (15, 18)
$\mathbf{F}_2$: Mental health status		
Distress level (Dis)	Nonspecific psychological distress scale (K6) score	2.01 (0, 2)
Impairment (Imp)	Daily functional impairment due to problems with emotions, nerves, or mental health	1.09 (0, 3)
Suicidal thought (Sui)	Serious thoughts of suicide in the past year (Yes=1/No=0)	%Yes: 9.58
Depression (Dep)	Major depressive episode in the past year (Y=1/N=0)	%Yes: 12.5
$\mathbf{F}_3$: Socioeconomic status		
Education (Edu)	5th grade or less (=5), 6th grade (=6), …, Freshman/13th year (=13), Sophomore/Junior (=14), Senior/Grad or more (=15)	12.41 (12, 14)
Insurance (Ins)	Having any health insurance (Y/N)	%Yes: 71.75
Family income (Fam)	Less than \$10,000 (=1), ~\$19,999 (=2), ~\$29,999 (=3), …, ~\$39,999 (=4), ~\$49,999 (=5), …, ~\$74,999 (=6), \$75,000 or more (=7)	4 (2, 6)
Employment Status (Emp)	Employed (Y=1/N=0)	%Yes: 67.02
Partitioning Covariates		
Age [b]	Groups of 18YearsOld, 19YO, 20YO, 21YO, 22/23YO, 24/25YO, b/w26-29YO, b/w30-34YO, b/w35-49YO, b/w50-64YO, or 65YO-older	27.38 (21, 32)
Gender	Male / Female	%Male: 54.64
Marital status (been married)	Married (N=1,797), Widowed (=83), Divorced/Separated (=1,072), Single/never been married (=5,460)	-
Ethnicity	Non-Hispanic-White, Hispanic, Non-Hispanic-All [c]	%:68.93/11.73/19.34

[a] For continuous variables, the first quartile (Q1), mean, and third quartile (Q3) are given.
[b] In the original survey, the age of each respondent was encoded as an ordinal variable. The group of 22/23 years old is the most dominant one, 17.27%. The average % of the other age groups are 9.09%.
[c] The category of "Non-Hispanic-All" includes non-Hispanic Native American/Alaskan Natives, non-Hispanic Hawaiians/other Pacific Islanders, non-Hispanic Asians, and people reporting more than one race (other than Hispanic).

Table 3. A summary of the parameter instability tests for the 2012 NSDUH data. A test statistic value and *p*-value are given for each partitioning covariate. The node numbers are consistent with those in Figure 4.

		Age		Gender		Marital Status		Ethnicity	
	Node	Statistic	*p*-value	Statistic	*p*-value	Statistic	*p*-value	Statistic	*p*-value
(a) Training set	1	18.57	.02	2.15	.96	12.73	.54	7.52	.00
	3	24.78	.00	.45	1.00	16.42	.17	0 [a]	-
(b) Test Set	1	27.99	.00	3.18	.84	19.27	.09	44.01	.00
	3	31.84	.00	.98	.99	21.58	.03	0 [a]	-

[a] Node 3 is ethnically homogeneous.

Table 4. The component weight estimates (Est.), and their standard errors (S.E.) and *p*-values from MOB-ERA for the 2012 NSDUH data.

			(a) Training set			(b) Test set		
Subgroups	Components	Predictors	Est.	S.E.	*p*-val	Est.	S.E.	*p*-val
Node 2	$\mathbf{F}_1$	Cigarette initiation (w_{11})	**.87**	.30	.00	**.97**	.29	.00
		Alcohol initiation (w_{12})	-.26	.30	.38	-.11	.29	.72
		Marijuana initiation (w_{13})	.39	.30	.19	.15	.30	.61
	$\mathbf{F}_2$	Distress level (w_{21})	.07	.29	.81	.38	.29	.20
		Impairment (w_{22})	**.50**	.25	.05	**.56**	.28	.05
		Suicidal thought (w_{23})	.31	.24	.20	-.06	.25	.81
		Depression (w_{23})	.45	.27	.10	.30	.27	.26
	$\mathbf{F}_3$	Education (w_{31})	**.75**	.24	.00	**.65**	.32	.04
		Insurance (w_{32})	**.52**	.23	.03	.17	.31	.58
		Family income (w_{33})	.12	.25	.62	.38	.32	.24
		Employment Status (w_{34})	.27	.24	.26	.57	.32	.07
Node 4	$\mathbf{F}_1$	Cigarette initiation (w_{11})	**.89**	.15	.00	**1.03**	.14	.00
		Alcohol initiation (w_{12})	**-.32**	.13	.02	**-.28**	.13	.03
		Marijuana initiation (w_{13})	**.33**	.15	.03	.27	.14	.05
	$\mathbf{F}_2$	Distress level (w_{21})	**.83**	.42	.05	**.67**	.34	.05
		Impairment (w_{22})	.23	.41	.57	.61	.33	.06
		Suicidal thought (w_{23})	.42	.34	.22	.02	.27	.93
		Depression (w_{23})	**-.68**	.38	.08	-.43	.29	.14
	$\mathbf{F}_3$	Education (w_{31})	**.87**	.11	.00	**.87**	.09	.00
		Insurance (w_{32})	**.33**	.10	.00	**.28**	.09	.00
		Family income (w_{33})	.09	.10	.37	.01	.09	.91
		Employment Status (w_{34})	.00	.10	1.00	.05	.09	.54
Node 5	$\mathbf{F}_1$	Cigarette initiation (w_{11})	**1.01**	.18	.00	**.67**	.21	.00
		Alcohol initiation (w_{12})	-.03	.18	.85	.04	.22	.84
		Marijuana initiation (w_{13})	-.13	.19	.47	**.53**	.22	.01
	$\mathbf{F}_2$	Distress level (w_{21})	.25	.20	.22	.46	.31	.14
		Impairment (w_{22})	.36	.20	.07	.13	.30	.66
		Suicidal thought (w_{23})	**.50**	.16	.00	.25	.23	.28
		Depression (w_{23})	.19	.18	.29	.43	.27	.11
	$\mathbf{F}_3$	Education (w_{31})	**.72**	.10	.00	**.55**	.11	.00
		Insurance (w_{32})	.15	.09	.11	.17	.11	.12
		Family income (w_{33})	.11	.10	.25	**.41**	.12	.00
		Employment Status (w_{34})	**.45**	.10	.00	**.34**	.11	.00

Table 5. The regression coefficient estimates (Est.), and their standard errors (S.E.) and *p*-values from MOB-ERA for the 2012 NSDUH data.

		$\mathbf{F}_1$: Substance initiation			$\mathbf{F}_2$: Mental health status			$\mathbf{F}_3$: Socioeconomic status		
	Node	Est.	S.E.	*p*-val	Est.	S.E.	*p*-val	Est.	S.E.	*p*-val
(a) Training set	2 (*N*=1,298)	-.11	.03	.00	.12	.03	.00	-.12	.03	.00
	4 (*N*=1,609)	-.23	.03	.00	.09	.03	.00	-.28	.03	.00
	5 (*N*=1,299)	-.14	.02	.00	.16	.02	.00	-.26	.02	.00
(b) Test Set	2 (*N*=1,316)	-.11	.03	.00	.12	.03	.00	-.09	.03	.00
	4 (*N*=1,591)	-.22	.03	.00	.10	.02	.00	-.28	.03	.00
	5 (*N*=1,299)	-.14	.03	.00	.13	.03	.00	-.25	.03	.00

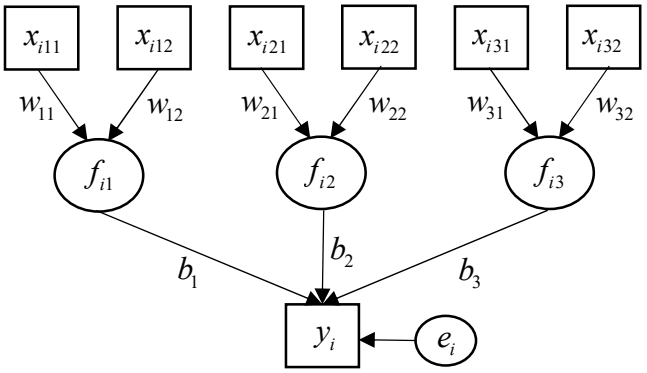


Figure 1. An exemplary ERA model. Square boxes indicate observed predictor and response variables. Circles represent components and an error term.

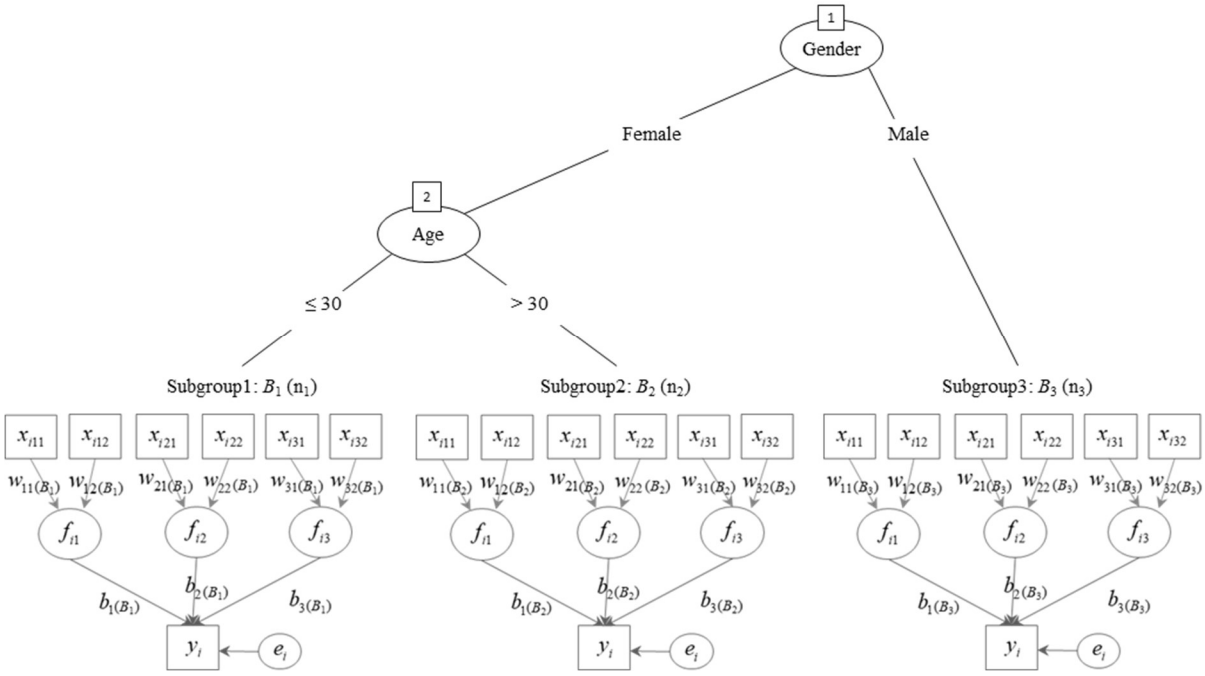


Figure 2. An illustrative example of MOB-ERA. Gender and age are used as partitioning covariates. Each identified subgroup provides its own ERA parameter estimates.

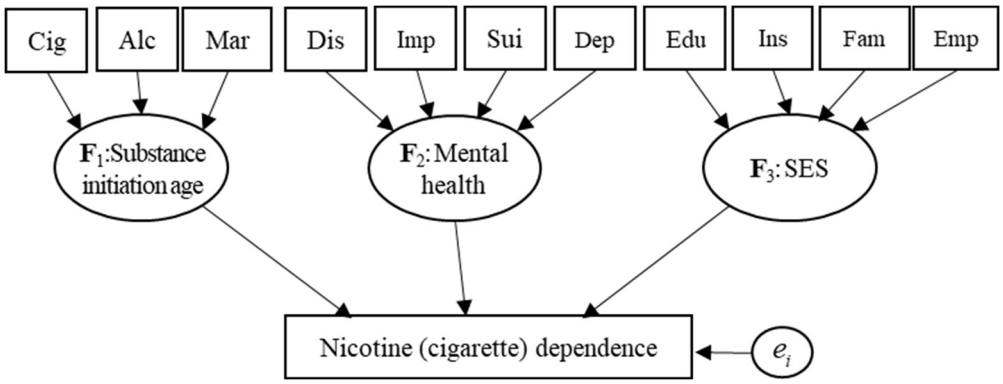
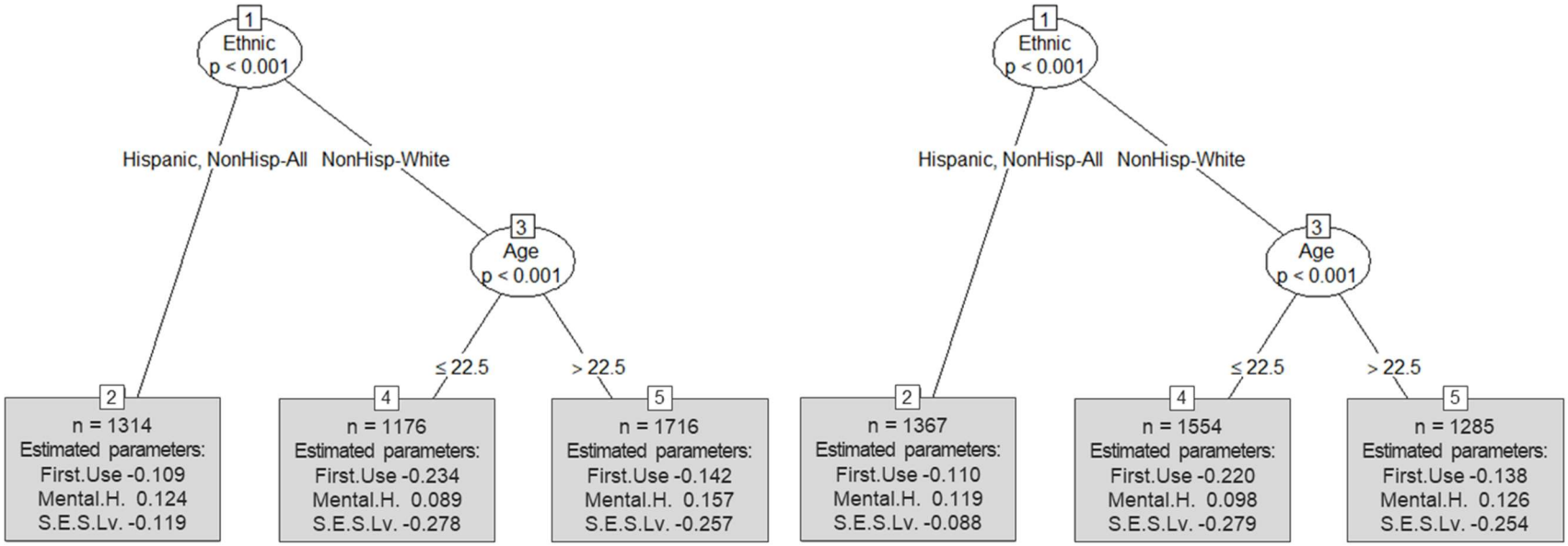


Figure 3. The ERA model for the 2012 NSDUH data. Variable names are consistent with those in Table 2.



(a) Training set (N_{train} = 4,206) (b) Test set (N_{test} = 4,206)

Figure 4. MOB-ERA for the 2012 NSDUH data: The final MOB-ERA trees obtained from (a) the training set and (b) the test set. Node numbers are given at the top of every internal (circle) and terminal (grey box) node. Each internal node corresponds to the selected partitioning covariate and its *p*-value obtained from the parameter instability test. Each terminal node represents an identified subgroup and provides the number of observations and regression coefficient estimates for the subgroup.

Appendix: Estimation and Inference in ERA

We express (7) in matrix notation as

$$\varphi = (\boldsymbol{z} - \mathbf{X}\mathbf{W}\boldsymbol{b})'\boldsymbol{\Omega}(\boldsymbol{z} - \mathbf{X}\mathbf{W}\boldsymbol{b}) = (\boldsymbol{z} - \mathbf{F}\boldsymbol{b})'\boldsymbol{\Omega}(\boldsymbol{z} - \mathbf{F}\boldsymbol{b}) \tag{A1}$$

with respect to $\mathbf{W}$ and $\boldsymbol{b}$, subject to $\text{diag}(\mathbf{F}'\mathbf{F}) = N\mathbf{I}$, where $\boldsymbol{z}$ is an N by 1 vector of adjusted response variable values z_i, $\mathbf{X}$ is an N by P matrix of predictors, $\mathbf{W}$ is a P by K matrix of component weights, $\boldsymbol{b}$ is a K by 1 vector of regression coefficients, $\boldsymbol{\Omega}$ is an N by N diagonal matrix of the ith diagonal element ω_i, and $\mathbf{F}$ is an N by K matrix of component scores.

To estimate ERA parameters, we aim to minimize (A1) by an iterative method in which each iteration involves the following steps:

Step1. Update $\mathbf{W}$ for fixed $\boldsymbol{b}$, $\boldsymbol{z}$, and $\boldsymbol{\Omega}$. This is equivalent to minimizing the following criterion with respect to $\mathbf{W}$,

$$\begin{aligned}
\varphi_{(\mathbf{W})} &= (\boldsymbol{z} - \mathbf{X}\mathbf{W}\boldsymbol{b})'\boldsymbol{\Omega}(\boldsymbol{z} - \mathbf{X}\mathbf{W}\boldsymbol{b}) \\
&= [\text{vec}(\boldsymbol{z} - \mathbf{X}\mathbf{W}\boldsymbol{b})]'\boldsymbol{\Omega}[\text{vec}(\boldsymbol{z} - \mathbf{X}\mathbf{W}\boldsymbol{b})] \\
&= [\boldsymbol{z} - (\boldsymbol{b}' \otimes \mathbf{X})\text{vec}(\mathbf{W})]'\boldsymbol{\Omega}[\boldsymbol{z} - (\boldsymbol{b}' \otimes \mathbf{X})\text{vec}(\mathbf{W})] \\
&= (\boldsymbol{z} - \mathbf{U}\boldsymbol{w}^*)'\boldsymbol{\Omega}(\boldsymbol{z} - \mathbf{U}\boldsymbol{w}^*)
\end{aligned} \tag{A2}$$

where $\otimes$ indicates the Kronecker product, vec($\mathbf{W}$) indicates the vec operator that creates the column vector of $\mathbf{W}$ obtained by stacking the columns of $\mathbf{W}$, $\mathbf{U}$ denotes an N by P matrix formed by eliminating the columns of $\boldsymbol{b}' \otimes \mathbf{X}$ corresponding to the nonzero elements in vec($\mathbf{W}$), and $\boldsymbol{w}^*$ denotes the P by 1 vector of the nonzero elements in vec($\mathbf{W}$). Then, the estimates of $\boldsymbol{w}^*$ are obtained by

$$\hat{\boldsymbol{w}}^* = (\mathbf{U}'\boldsymbol{\Omega}\mathbf{U})^{-1}\mathbf{U}'\boldsymbol{\Omega}\boldsymbol{z}\,. \tag{A3}$$

Subsequently, the nonzero elements in $\mathbf{W}$ are replaced with the corresponding values in $\boldsymbol{w}^*$.

Step2. Update $\boldsymbol{b}$ for fixed $\mathbf{W}$, $\boldsymbol{z}$, and $\boldsymbol{\Omega}$. This is equivalent to minimizing

$$\begin{aligned}\varphi_{(\mathbf{b})} &= (\boldsymbol{z} - \mathbf{XW}\boldsymbol{b})'\boldsymbol{\Omega}(\boldsymbol{z} - \mathbf{XW}\boldsymbol{b}) \\ &= (\boldsymbol{z} - \mathbf{F}\boldsymbol{b})'\boldsymbol{\Omega}(\boldsymbol{z} - \mathbf{F}\boldsymbol{b})\end{aligned} \tag{A4}$$

with respect to $\boldsymbol{b}$, subject to diag($\mathbf{F}'\mathbf{F}$) = $N\mathbf{I}$. The least-squares estimate of $\boldsymbol{b}$ is given by

$$\hat{\boldsymbol{b}} = (\mathbf{F}'\boldsymbol{\Omega}\mathbf{F})^{-1}\mathbf{F}'\boldsymbol{\Omega}\boldsymbol{z}\,. \tag{A5}$$

Step3. Update $\boldsymbol{z}$ and $\boldsymbol{\Omega}$ for fixed $\mathbf{W}$ and $\boldsymbol{b}$. As discussed in the Methods section, $\boldsymbol{z}$ is updated based on $z_i = \eta_i + (y_i - \mu_i)/\omega_i$. The calculation of $\boldsymbol{\Omega}$ varies depending on which member of the exponential family is assumed for the response variable (refer to McCullagh & Nelder, 1989). For example, in the case of the normal distribution, $\hat{\omega}_i = \hat{\mu}_i^{-2} = 1$ yielding $\boldsymbol{\Omega} = \mathbf{I}_N$.

We repeat the above steps until the changes in $\hat{\mathbf{W}}$ and $\hat{\boldsymbol{b}}$ between previous and current iterations are below a pre-determined threshold, e.g., 10^{-5}.

Let $\hat{\boldsymbol{\theta}}_{\text{ERA}} = [\,\hat{\boldsymbol{w}}^*; \hat{\boldsymbol{b}}\,]$ denotes the ML parameter estimates at convergence that stacks $\hat{\boldsymbol{w}}^*$ and $\hat{\boldsymbol{b}}$. The asymptotic covariance matrix of $\hat{\boldsymbol{\theta}}_{\text{ERA}}$ can be obtained by computing negative Hessian matrix evaluated at $\hat{\boldsymbol{\theta}}_{\text{ERA}}$ and inverting it (Hwang et al., 2015; Yee & Hastie, 2003). Let $\hat{\boldsymbol{\theta}}_{\text{ERA}} = \hat{\boldsymbol{\theta}}$ for simplicity. The negative Hessian matrix or the second-derivative of the log-likelihood is given as

$$-\mathrm{H}(\boldsymbol{\theta}) = -\frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'} = -\begin{pmatrix} \dfrac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \boldsymbol{w}^* \partial \boldsymbol{w}^{*\prime}} & \dfrac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \boldsymbol{w}^* \partial \boldsymbol{b}'} \\ \dfrac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \boldsymbol{b} \partial \boldsymbol{w}^{*\prime}} & \dfrac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \boldsymbol{b} \partial \boldsymbol{b}'} \end{pmatrix}. \tag{A6}$$

The diagonal terms in (A6) can be obtained by fixing $\boldsymbol{w}^*$ and $\boldsymbol{b}$, respectively:

$$-\frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \boldsymbol{w}^* \partial \boldsymbol{w}^{*\prime}} = -\mathbf{U}'\boldsymbol{\Omega}\mathbf{U} \tag{A7}$$

and

$$-\frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \boldsymbol{b} \partial \boldsymbol{b}'} = \mathbf{F}'\boldsymbol{\Omega}\mathbf{F}. \tag{A8}$$

The off-diagonal terms can be obtained using the profile likelihoods (Richards, 1961)

$$-\frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \boldsymbol{b} \partial \boldsymbol{w}^{*\prime}} = -\frac{\partial \boldsymbol{w}^{*\prime}}{\partial \boldsymbol{b}}\left(-\frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \boldsymbol{w}^{*} \partial \boldsymbol{w}^{*\prime}}\right). \tag{A9}$$

To compute $-\frac{\partial \boldsymbol{w}^{*\prime}}{\partial \boldsymbol{b}}$ in (A8), let $\boldsymbol{\delta}_j$ denote a K by 1 vector of 0 except having 1 in the jth element ($j = 1, \ldots, K$) and Λ denote a matrix formed by eliminating the columns of $\boldsymbol{\delta}'_j \otimes \mathbf{X}$ corresponding to the fixed elements in vec($\mathbf{W}$). Then, $-\frac{\partial \boldsymbol{w}^{*\prime}}{\partial \boldsymbol{b}}$ is calculated by

$$-\frac{\partial \boldsymbol{w}^{*\prime}}{\partial \boldsymbol{b}} = (\mathbf{U}'\boldsymbol{\Omega}\mathbf{U})^{-1}\left[\Lambda'\boldsymbol{\Omega}\boldsymbol{z} - \Lambda'\boldsymbol{\Omega}\mathbf{U}\left((\mathbf{U}'\boldsymbol{\Omega}\mathbf{U})^{-1}\mathbf{U}'\boldsymbol{\Omega}\boldsymbol{z}\right)\right] \quad (j = 1, \ldots, K). \tag{A10}$$