Interpretable Data Reduction in Prediction Modeling:

Extended Redundancy Analysis and Its Extensions and Applications

Sunmee Kim

Department of Psychology

McGill University, Montreal, Canada

April 2020

A dissertation submitted to McGill University in partial fulfillment of

the requirements of the degree of Doctor of Philosophy.

Table of Contents

Abstract

With technical advances in measuring human behaviors and psychological traits, research goals in psychology increasingly call for statistical methods for data with complex structure. Modeling such data is challenging because they typically contain a large number of redundant predictor dimensions and potentially heterogeneous subgroups of observations. Moreover, properly handling the covariance structure of repeated or clustered measures is critical when observations are correlated rather than independent. Assessment of model performance on independent unseen data is also important as it guides the choice of final model structure when researchers seek to develop models that can generalize beyond the current sample.

The present research proposes solutions to these specific challenges in the framework of extended redundancy analysis (ERA). ERA is a statistical tool that performs data reduction and regression analysis simultaneously. When investigating a complex social and behavioral phenomenon that involves multiple sets of numerous predictors, ERA can be useful as it provides a comprehensible description of predictor-response relationships by summarizing each set of predictors into its low-dimensional representation—a component. However, conventional ERA has no efficient mechanism to account for unknown but potential heterogeneity in data. Also, it is unable to support analysis of multiple correlated responses without assuming a multivariate normal distribution. Furthermore, all existing model evaluation metrics for ERA are “in-sample” metrics, which may limit the generalizability of the selected model.

To address these challenges, this dissertation presents two novel extensions of ERA and suggests an alternative perspective of model assessment. The first extension combines ERA with a recursive partitioning method to automatically identify heterogeneous subgroups

of observations differentiated by auxiliary covariates (e.g., gender, ethnicity, etc.). The second extension adopts generalized estimating equations in order to model correlated response variables with an unknown covariance structure. Both ERA extensions adopt various regularization techniques, such as pruning or $L2$ regularization, thus being able to avoid overfitting and determine the model complexity in a data-driven manner. The theoretical underpinnings of the two proposed methods are discussed in detail, along with an illustration of their empirical usefulness using data from the 2012 National Survey on Drug Use and Health in the US. Finally, this dissertation also demonstrates the benefit of using various resampling methods to advance the existing in-sample model assessment in ERA based on the analyses of simulated data.

Abrégé

Avec les progrès techniques dans la mesure des comportements humains et des traits psychologiques, les objectifs de la recherche en psychologie exigent de plus en plus des méthodes statistiques pour les données à structure complexe. La modélisation de ces données est difficile car elles contiennent généralement un grand nombre de dimensions non pertinentes ou redondantes et des sous-groupes d'observations potentiellement hétérogènes. De plus, la gestion appropriée de la structure de covariance des mesures répétées ou groupées est essentielle lorsque les observations sont corrélées plutôt qu'indépendantes. L'évaluation des performances du modèle sur des données indépendantes invisibles est également importante car elle guide le choix de la structure finale du modèle lorsque les chercheurs cherchent à développer des modèles qui peuvent généraliser au-delà de l'échantillon actuel.

La présente recherche propose des solutions à ces défis spécifiques dans le cadre de l'analyse de redondance étendue (extended redundancy analysis; ERA). L'ERA est un outil statistique qui effectue simultanément une réduction des données et une analyse de régression. Lorsque vous étudiez un phénomène social et comportemental complexe qui implique plusieurs ensembles de nombreux prédicteurs, l'ERA peut être utile car elle fournit une description compréhensible des relations prédicteur-réponse en résumant chaque ensemble de prédicteurs dans sa représentation de faible dimension - un composant. Cependant, l'ERA conventionnelle n'a pas de mécanisme efficace pour tenir compte de l'hétérogénéité inconnue mais potentielle des données. En outre, il n'est pas en mesure de soutenir l'analyse de réponses corrélées multiples sans supposer une distribution normale multivariée. En outre, toutes les mesures d'évaluation de modèle existantes pour l'ERA sont des mesures «dans l'échantillon», ce qui peut limiter la généralisabilité du modèle sélectionné.

Pour relever ces défis, cette thèse présente deux nouvelles extensions de l'ERA et suggère une perspective alternative de l'évaluation du modèle. La première extension combine l'ERA avec une méthode de partitionnement récursif pour identifier des sous-groupes hétérogènes d'observations différenciées par des covariates auxiliaires (par exemple, le sexe, l'origine ethnique, etc.). La deuxième extension adopte des équations d'estimation généralisées afin de modéliser des variables de réponse corrélées avec une structure de covariance inconnue. Les deux extensions ERA adoptent diverses techniques de régularisation, telles que l'élagage ou la régularisation *L2*, permettant ainsi d'éviter le sur-ajustement et de déterminer la complexité du modèle en fonction des données. Les fondements théoriques des deux méthodes proposées sont examinés en détail, ainsi qu'une illustration de leur utilité empirique à l'aide des données de l'enquête nationale de 2012 sur la consommation de drogues et la santé aux États-Unis. Enfin, cette thèse démontre également l'avantage d'utiliser diverses méthodes de rééchantillonnage pour faire progresser l'évaluation du modèle existant dans l'échantillon en ERA sur la base des analyses de données simulées.

Acknowledgements

I am most grateful to my supervisor, Professor Heungsun Hwang, who has so generously offered his expertise as I work on reaching my academic goal. While undertaking this Ph.D. degree, I have been learning, growing, and becoming a better version of myself as a researcher thanks to his guidance and inspiration. His mentorship has motivated me to do better, and his feedback on my performance has helped me strengthen all of my skills. I admire his intelligence and dedication at work, and I am truly fortunate to have worked with him.

I would like to thank a number of faculty members in the Department of Psychology at McGill University. I gratefully acknowledge the advice and constructive feedback provided by Professor Carl F. Falk, Jessica Kay Flake, and Milica Miočević. Their words of encouragement have a much bigger impact than they will ever know. I am genuinely glad that I could spend my final years of graduate work with them. Past and current colleagues of my research—Ramsey Cardwell, Professor Ji-Yeh Choi, Dr. Sungyoung Lee, and Gyeongcheol Cho—who contributed their time and ideas, deserve a special thanks, as well.

All my success and achievements, I owe to my family. They have been amazingly understanding as I navigate this life-changing experience—I am so lucky to have them in my corner. Without my parents' endless love and care, I don't know where I would be. My sister, Seon-ok, my best friend and soulmate, her thoughtfulness and generosity empower me and encourage me to do my best, always. Needless to say, I am indebted to my dearest friends, Shawn Suyong Yi Jones, Amélie Bernard, and Robert Margaryan who have supported me no matter what—they have been my lifeline in Canada.

Lastly, I want to take this opportunity to express my gratitude to McGill's faculty and staff members: Although we are still dealing with much change given the COVID-19

outbreak, they have been working hard to give guidance and adjust to the uncertainties during this extraordinary time. I would like to extend to them my sincere thanks well wishes. I hope the situation clears up soon enough that I will have a chance to get together with my family, close friends, and academic advisors before I move on to my next career adventure.

Contribution of Authors

This dissertation is presented as a manuscript-based thesis based on the text of one manuscript submitted for publication, one in press, and one published which together present a unified research project, novel technical extensions of conventional ERA. These manuscripts were presented in international conferences as well. The full bibliographic information of manuscripts and conference presentations is as follows:

- Chapter 2. Model-Based Recursive Partitioning of ERA
    - Manuscript [1]: **Kim, S.** & Hwang, H. Model-Based Recursive Partitioning of Extended Redundancy Analysis with an Application to Nicotine Dependence among US adults. Submitted to *British Journal of Mathematical and Statistical Psychology* (2019).
    - Conference Presentation: **Kim, S.** & Hwang, H. (July 2017). *Recursive Partitioning of Extended Redundancy Analysis.* Presented at the International Meeting of the Psychometric Society, Zurich, Switzerland.
- Chapter 3. Regularized ERA via Generalized Estimating Equations
    - Publication [2]: **Kim, S.***, Lee, S.*, Cardwell, R., Kim, Y., Park, T., & Hwang, H. (in press, May 2020). An application of regularized extended redundancy analysis via generalized estimating equations to the study of co-occurring substance use among US adults. *Quantitative Psychology. IMPS 2019. *Co-first author*
    - Publication [3]: 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,* Article number: 100. https://doi.org/10.1186/s12920-019-0517-4. *Co-first author*
    - Conference Presentation: **Kim, S.** & Hwang, H. (July 2019). *Extended Redundancy Analysis via Generalized Estimating Equations.* Presented at the International Meeting of the Psychometric Society, Santiago, Chile.

In the dissertation, the submitted manuscript [1] and the book chapter in press [2] are reproduced in whole in Chapter 2 and Chapter 3, respectively. The article [3] published in a peer-reviewed journal in genomics is included in part, with the methodology appearing in Appendix C. Parameter Estimation in GEE-ERA, to make the present research coherent.

I am the first author of the submitted manuscript [1], the co-first and corresponding author of [2], and the co-first author of [3]. My doctoral supervisor Professor Heungsun Hwang provided important feedback, guidance, and inspirations at every stage of all the research. I acknowledge the constructive feedback of Ramsey Cardwell (Ph.D. Candidate, Quantitative Psychology, University of North Carolina at Greensboro, NC, US) in the further improvements of the paper [2]. The paper [3] is co-authored by Dr. Sungyoung Lee (Seoul National University Hospital, Seoul, South Korea) recognizing his contribution in the analysis of large-scale whole-exome sequencing data, simulation studies of rare genetic variants, and writing of the related sections. Beyond the noted contributions, the whole work presented in this dissertation, including the methodological contributions of the main algorithms, data processing and analysis, simulation studies, and writing of the text, was completed and written by the author.

List of Tables

List of Figures

Chapter 1. Introduction and Background

1.1. Extended Redundancy Analysis

Extended Redundancy Analysis (ERA; Takane & Hwang, 2005) is a statistical modeling framework that performs dimension reduction and regression analysis simultaneously to investigate the directional relationships between multiple sets of predictors and response variables. In ERA, each set of predictors is summarized into its low-dimensional representation—a component, which in turn predicts response variables. A set of predictors related to a component can be selected based on prior theories or domain-specific knowledge to facilitate the interpretability of the component. 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), multivariate ERA for response variables from a multivariate normal distribution (Lee, Kim, Choi, Hwang, & Park, 2018), functional ERA for the analysis of smooth functions or curves (Hwang, Suk, Takane, Lee, & Lim, 2015; Tan, Choi, & Hwang, 2015), and Bayesian ERA (Choi, Kyung, Hwang, & Park, 2020).

Linear regression, one of the most common statistical methods applied in various psychological studies to provide an interpretable description of how predictors affect response variable(s), assumes there are no strong correlations among any subsets of the predictors entered in a model, or no multicollinearity. Unfortunately, a regression model with a large pool of potential predictors rarely meets this stringent assumption of no multicollinearity (Cheung & Jak, 2016; Farrar & Glauber, 1967; Hastie, Tibshirani, & Friedman, 2009, Chapter 3; Moustafa et al., 2018; Smith & Sasaki, 1979). For example, studies on recreational psychoactive substance use among American adults have

demonstrated various predictors that are possibly highly correlated (Daza et al., 2006; Hu, Davies, & Kandel, 2006a; Kandel, Kiros, Schaffran, & Hu, 2004; Robinson et al., 2006). Such predictors include initiation age of specific substances (e.g., cigarette, alcohol, and/or marijuana), indicators of mental health or mental disorder, and variables concerning socioeconomic status (SES), to name a few. Regression analysis in this setting is not satisfactory because it can suffer from potential multicollinearity problems and, moreover, is incapable of providing a comprehensible description of directional relationships between many sets of predictors and response variables. Other widely noted examples of data with highly correlated predictors include those used in large-scale neuroimaging or genetic studies. Data collected in such studies typically include hundreds or thousands of brain voxel-level phenotypes (e.g., blood-oxygen-level-dependent or BOLD time series) or nucleotide-level variants (e.g., single nucleotide polymorphisms or SNPs). The brain phenotypes or genetic variants are grouped into particular brain or genomic locations, and for example, thousands of neighboring SNPs at a genomic location are often highly correlated (Day et al., 2015; Gusev et al., 2016; Kozberg, Chen, DeLeo, Bouchard, & Hillman, 2013; Spano et al., 2013).

For such studies, there are both statistical and practical benefits of using ERA, compared to using conventional regression analysis. Statistically, ERA circumvents possible multicollinearity problems by replacing a large number of original predictors with a (much) smaller number of uncorrelated components for regression-based prediction. Such data reduction procedure in ERA ensures the predictability of the final model as well, because each component is extracted from each set of predictors in such a way that it accounts for the maximum variation of the responses. This is in fact a win-win strategy, statistically and practically, because ERA provides a parsimonious and interpretable solution to a regression problem that involves a rich set of psychological, physiological, and socio-demographic

predictors using the extracted components, whose number is much smaller than that of the original predictors. Using domain-specific knowledge of which predictors can be put together to form a component improves the interpretability of components.

1.2. Model Specification and Statistical Inference

In this section, the ERA model and its statistical inference are recapitulated to facilitate an understanding of the relevant methods and the new extensions that will be discussed in the following sections and chapters.

Let y_{iq} denote the ith value of the qth response variable ($i = 1, \cdots, N$; $q = 1, \cdots, Q$). Assume that there are K different sets of predictors, each of which consists of P_k predictors ($k = 1, \cdots, K$). Let x_{ikp} denote the ith value of the pth variable in the kth predictor set ($p = 1, \cdots, P_k$) 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 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_{kq} denote the regression coefficient relating the kth component to the qth response variable. Let e_{iq} denote an error term for y_{iq}. We assume that all the predictors are standardized with zero means and unit variances (Takane & Hwang, 2005).

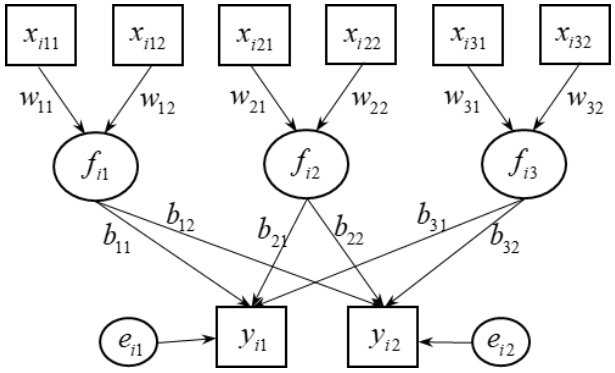


Figure 1.2.1. A prototypical ERA model

The ERA model (Takane & Hwang, 2005) is then expressed as

$$\begin{aligned} y_{iq} &= \sum_{k=1}^{K}\left[\sum_{p=1}^{P_k} x_{ikp} w_{kp}\right] b_{kq} + e_{iq} = \boldsymbol{x}_i'\mathbf{W}\boldsymbol{b}_q + e_{iq} \\ &= \sum_{k=1}^{K} f_{ik} b_{kq} + e_{iq} = \boldsymbol{f}_i\boldsymbol{b}_q + e_{iq}\ , \end{aligned} \tag{1.2.1}$$

where $\mathbf{W} = \text{diag}(\boldsymbol{w}_1,\cdots,\boldsymbol{w}_K)$, $\boldsymbol{f}_i' = (f_{i1},\cdots,f_{iK})$, and $\boldsymbol{b}_q = (b_{1q},\cdots,b_{Kq})'$. This can also be expressed in matrix notation as

$$\begin{aligned} \mathbf{Y} &= \mathbf{XWB} + \mathbf{E} \\ &= \mathbf{FB} + \mathbf{E}, \end{aligned} \tag{1.2.2}$$

where **Y** is an *N* by *Q* matrix of response variables, **X** is an *N* by *P* matrix of predictors, **W** is a *P* by *K* matrix of weights, **B** is a *K* by *Q* matrix of regression coefficients, and **E** is an *N* by *Q* matrix of errors. For identification of **F**, a standardization constraint is imposed on **F** such that diag(**F**′**F**)=*N***I**. As shown in (1.2.1), each set of predictors reduces to a single component, which in turn influences the *q*th response variable. The component weight w_{kp} shows the contribution of each predictor to obtaining its component as in data reduction methods such as principal component analysis or canonical correlation analysis, whereas the regression coefficient b_{kq} signifies the effect of each component on each response variable as in linear

regression. In this regard, ERA carries out data reduction and linear regression simultaneously, as discussed earlier.

Figure 1.2.1 displays an example of the ERA model, where each component is associated with two predictors ($P_1 = P_2 = P_3 = 2$) and two response variables ($Q = 2$) are influenced by three components ($K = 3$). In the figure, squares indicate predictors and response variables, whereas circles represent components or error terms. For this example, the **W** and **B** matrices are given as

$$\mathbf{W} = \text{diag}(\boldsymbol{w}_1, \boldsymbol{w}_2, \boldsymbol{w}_3) = \begin{pmatrix} w_{11} & 0 & 0 \\ w_{12} & 0 & 0 \\ 0 & w_{21} & 0 \\ 0 & w_{22} & 0 \\ 0 & 0 & w_{31} \\ 0 & 0 & w_{32} \end{pmatrix} \quad \text{and} \quad \mathbf{B} = \begin{pmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \\ b_{31} & b_{32} \end{pmatrix}.$$

The ERA model contains two sets of parameters to be estimated—component weights (**W**) and regression coefficients (**B**). These unknown parameters are estimated by minimizing the following least-squares objective function:

$$\phi = \text{SS}\left(\mathbf{Y} - \mathbf{XWB}\right), \tag{1.2.3}$$

with respect to **W** and **B**, subject to the constraint diag(**F**'**F**)=*N***I**, where SS(**A**)=trace(**A**'**A**). An alternating least-squares (ALS) algorithm (de Leeuw, Young, & Takane, 1976) was developed to minimize the objective function (Takane & Hwang, 2005). The ALS algorithm alternates two main steps until convergence:

Step 1. Update **W** for fixed **B**. This is equivalent to minimizing the following criterion with respect to **W**,

$$\begin{aligned} \phi &= \text{SS}\left(\text{vec}\left(\mathbf{Y}\right) - \text{vec}\left(\mathbf{XWB}\right)\right) \\ &= \text{SS}\left(\text{vec}\left(\mathbf{Y}\right) - \left(\mathbf{B}' \otimes \mathbf{X}\right)\text{vec}\left(\mathbf{W}\right)\right), \end{aligned} \tag{1.2.4}$$

where vec(**A**) indicates the vec operator that creates the column vector of a matrix **A** obtained by stacking the columns of **A** and $\otimes$ refers to the Kronecker product. Let $\mathbf{W}^*$ denote a column vector formed by eliminating zero elements from vec(**W**), and **Ω** denote a matrix formed by eliminating the columns of $\mathbf{B}' \otimes \mathbf{X}$ corresponding to the zero elements in vec(**W**). The least-squares estimate of $\mathbf{W}^*$ is then obtained by

$$\hat{\mathbf{W}}^* = \left(\mathbf{\Omega}'\mathbf{\Omega}\right)^{-1}\mathbf{\Omega}'\text{vec}\left(\mathbf{Y}\right). \tag{1.2.5}$$

Subsequently, the nonzero elements in **W** are replaced with the corresponding values in $\hat{\mathbf{W}}^*$. Then, the updated **W** is multiplied by $\sqrt{N(\mathbf{W}'\mathbf{X}'\mathbf{X}\mathbf{W})^{-1}}$ to satisfy the constraint diag(**F'F**)=*N***I**.

Step 2. Update **B** for fixed **W**. This is equivalent to minimizing

$$\begin{aligned}\phi &= \text{SS}\left(\text{vec}\left(\mathbf{Y}\right) - \text{vec}\left(\mathbf{XWB}\right)\right)\\ &= \text{SS}\left(\text{vec}\left(\mathbf{Y}\right) - \left(\mathbf{I}\otimes\mathbf{F}\right)\text{vec}\left(\mathbf{B}\right)\right)\\ &= \text{SS}\left(\text{vec}\left(\mathbf{Y}\right) - \mathbf{\Gamma}\mathbf{B}^*\right),\end{aligned} \tag{1.2.6}$$

where **Γ** is a matrix formed by removing the columns in $\mathbf{I}\otimes\mathbf{F}$ that correspond to zero elements in vec(**B**), and $\mathbf{B}^*$ is a column vector formed by eliminating zero elements from vec(**B**). The least-squares solution for $\mathbf{B}^*$ is then given by

$$\hat{\mathbf{B}}^* = \left(\mathbf{\Gamma}'\mathbf{\Gamma}\right)^{-1}\mathbf{\Gamma}'\text{vec}\left(\mathbf{Y}\right). \tag{1.2.7}$$

Similarly, the updated **B** is reconstructed by $\hat{\mathbf{B}}^*$.

To assess the performance of ERA models, an overall fit measure, called FIT (Takane & Hwang, 2005), can be calculated to evaluate how well a given ERA model explains the variance of the response variables:

$$\text{FIT} = 1 - \frac{\phi}{SS(\mathbf{Y})}. \tag{1.2.8}$$

The values of FIT range from 0 to 1, and larger values indicate more explained variance of the variables[1].

To test the statistical significance of parameter estimates, ERA can use resampling methods, such as permutation tests for obtaining exact *p*-values (as described in Lee et al., 2016, 2018) and bootstrapping (Efron & Tibshirani, 1986) for constructing confidence intervals. For example, a 95% bootstrapped confidence interval, i.e., the 2.5th and 97.5th percentiles of the bootstrap distribution of a parameter estimate based on 1,000 bootstrapped replications, are often used. Although there is no strict rule of thumb for the number of bootstrap replications, in general, 500 to 1,000 bootstrap replications may be sufficient.

The ERA model described thus far is considered non-parametric because the distributions of response variables are unknown or not predetermined. More recent extensions of ERA can handle various types of response variables under some distributional assumptions. One advantage of having a distributional assumption is that standard errors, significance tests, or confidence intervals (CIs) are available without performing any resampling. For example, the constrained stochastic ERA model (DeSarbo, Hwang, Blank, & Kappe, 2015) is univariate (i.e., considers a single response variable), is linear in its ERA parameters, and has normally distributed residuals. The generalized ERA model proposed in Lee et al. (2016) assumes an exponential family distribution in order to accommodate phenotype data arising from a binomial distribution.[2] Also, in Lee et al. (2018), a multivariate normal distribution is assumed for the analysis of multiple response variables.

[1] As will be discussed in Chapter 4, other measures of overall model fit for parametric ERA are based on penalized-likelihood criteria, such as AIC_{ERA} and BIC_{ERA} (DeSarbo, Hwang, Blank, & Kappe, 2015), which take model complexity into account.

[2] Refer to the methodology in Chapter 2 (Appendix B) for the detailed description on the statistical inference of the generalized ERA model.

1.3. Review of Relevant Methods

Like ERA, principal components regression (PCR; Hotelling, 1957; Kendall, 1957) and partial least-squares regression (PLSR; de Jong, 1993; Wold, 1966) also incorporate data reduction into a regression problem. Thus, all three approaches are especially useful when regular regression produces unreliable coefficients with high standard errors or fails completely in the cases where predictors are highly collinear and/or the number of predictors is significantly larger than the number of observations. Technically, all three approaches aim to summarize the original predictors into a smaller set of uncorrelated components and perform least-squares regression on these components, where the components are defined as exact linear combinations of their associated predictors. In this view, they all attempt to capture most of information in **X** for predicting **Y** while reducing the dimensions of **X**. A major difference of the approaches arises from how components are constructed to explain response variables.

As illustrated in Figure 1.3.1, in PCR, principal component analysis (PCA) is first carried out to obtain principal components (PCs) of predictors, $PC_1, \dots, PC_P$, where P is the number of predictors. Often times, most of the variance in the original predictors can be captured by the first few principal components. Thus, the first k (for $k < P$) principal components, $PC_1, \dots, PC_k$, are then used to predict the response variable as in multiple regression (Hotelling, 1957; Jolliffe, 1982)[3]. When PCs are formed by spectral decomposition of the covariance matrix of predictors, they are extracted to maximize the variance of the predictors only, not considering how each predictor is related to the response variables. This implies that the selected PCs for regression may not be optimal in explaining

[3] Note that we get a reduced regression for $P < k$. Naturally, the problem of choosing an optimum subset of PCs remains. Typically, a scree plot that displays the eigenvalues of PCs (ordering the eigenvalues from largest to smallest) is used to determine the number of components to retain (k).

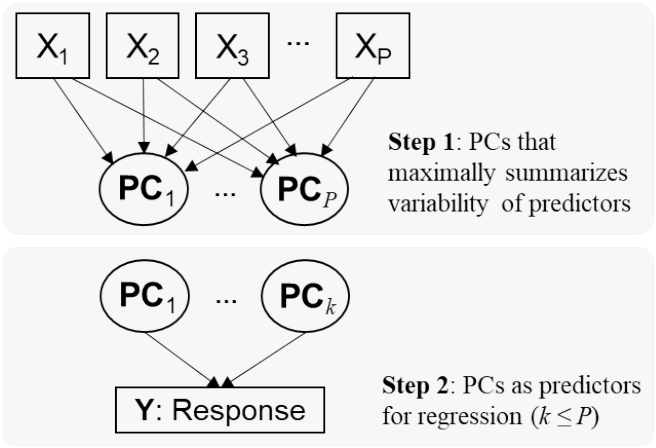


Figure 1.3.1. The steps involved in performing PCR

the variance of the response variables: e.g., the first PC, which accounts for the most variance in predictors, can be less explanatory for the response variable than the second or third PC (Maitra & Models, 2008, pp. 84-86). Some simulation studies show that even unselected PCs can explain a great deal of the variance in the response variable (e.g., Jolliffe, 1982; Smith & Campbell, 1980; Tian, Wilcox, & James, 2010). Moreover, when the covariance of **X** is characterized by only a few dominant eigenvalues and the number of non-zero eigenvalues is not negligible, PCR becomes less satisfactory because the PCs corresponding to small eigenvalues are all omitted (De Mol, Giannone, & Reichlin, 2008).

An alternative approach to PCR is PLSR, which also constructs components as a set of linear combinations of predictors (often called *latent vectors* in the PLS literature). But, unlike PCR, it takes into account the covariances between components and response variable in order to aid the construction of components that are maximally explain the variation of the response variable (Alin, 2009; Geladi & Kowalski, 1986; Mehmood & Ahmed, 2016). More specifically, PLSR searches for a set of components that performs a simultaneous decomposition of both **X** and **Y** with the constraint that these components explain the covariance between (the decompositions of) **X** and **Y** as much as possible (Refer to Appendix

A for details of these steps). To achieve this, PLSR uses an iterative algorithm, e.g., Nonlinear Iterative PArtial Least Squares (NIPALS; Wold, 1966) or Statistically Inspired Modification of the PLS method (SIMPLS; de Jong, 1993). Once the components are obtained, a regression step is followed to estimate the effects of the PLS components on response variables[4]. As such, the whole parameter estimation procedure of PLSR involves separate stages for estimating components and regression coefficients, optimizing multiple objective functions separately for each group of parameters. Such lack of a well-founded global optimization criterion makes it difficult to evaluate the overall goodness of fit of the specified model (Hwang & Takane, 2014; McDonald, 1996; Takane & Hwang, 2005).

Conversely, ERA employs an alternating least-squares algorithm to minimize the global least-squares criterion in (1.2.3) for parameter estimation. As discussed previously, this contributes to calculating a measure of overall model fit. Moreover, ERA considers multiple sets of predictors and reduces each set into a separate component (Figure 1.2.1), whereas PCR and PLSR extract more than one component from all the predictors entered in the analysis (Figure 1.3.1). Often, this makes it difficult to describe or understand the substantive meaning of obtained components for explaining response variables (Enki, Trendafilov, & Jolliffe, 2013; Shmueli, 2010), which in turn has limited the use of PCR and PLSR in fields such as psychology or medicine, where interpretation of "regressors" in the final model is important. Components that lack adequate substantive or theoretical grounds may contribute to the lack of interpretability of the final regression model. In ERA, however, each set of predictors falls into a distinct non-overlapping component, where each component

[4] As in PCR, the problem of determining the optimum number of PLS components in the final model remains, and cross-validation (CV) is widely used. For example, root mean squared errors (RMSE) are calculated for PLS models with increasing number of components, then the optimal number of components is chosen as the one that minimizes the RMSE. Similarly, the variable importance in the projection (VIP) method (Gauchi & Chagnon, 2001) and the Q^2 criterion (Stone, 1974) also utilize CV.

is well-defined on the basis of substantively meaningful specification of predictor group using priori domain-specific knowledge or assumptions about the data.

Researchers may attempt to accommodate an ERA model in the framework of structural equation modeling (SEM; Jöreskog, 1970, 1973), which is another widely used multivariate statistical method that examines the relationships between observed and latent variables. Conventional structural equation models, often referred to as factor-based SEM or covariance structure analysis in the SEM literature, assume that latent variables are equivalent to common factors which explain the covariance of observed variables only (Bollen, 1989; Bollen & Bauldry, 2011; Edwards & Bagozzi, 2000). On the other hand, ERA aims to obtain components as linear functions of predictors, which in turn explains the maximum variance of response variables, rather than common factors. Consequently, there are several technical difficulties when accommodating ERA components in the framework of SEM. For instance, including ERA components leads to an identification problem, violating the so-called 2+ emitted path rule (Bollen & Davis, 2009; MacCallum & Browne, 1993). Moreover, SEM is prone to the occurrence of non-convergence when the number of variables is large while the number of observations is limited (Bentler & Chou, 1987; Deng, Yang, & Marcoulides, 2018; Jackson, 2001). However, high-dimension low-sample-size is common in many social and behavioral studies, e.g., in the data collection of neuroimaging or genetic studies due to the high costs associated with obtaining a sufficient number of observations. As discussed previously, ERA has been adopted as a useful tool for fitting such high-dimensional data in various applications (e.g., DeSarbo et al., 2015; Hwang et al., 2015; Lee et al., 2016, 2018).

1.4. Dissertation Objectives and Overview

This dissertation is a manuscript-based thesis and presents two ERA extensions that I have recently proposed to address issues of substantive importance in psychology and related fields. Specifically, both extensions deal with how to effectively explore the relationships between substance use and various associated socio-psychological variables using data from the National Survey on Drug Use and Health (NSDUH; United States Department of Health and Human Services. Substance Abuse and Mental Health Services Administration. Center for Behavioral Health Statistics and Quality, 2013). As discussed earlier, the use of recreational psychoactive substances is an example of a complex social phenomenon that involves multiple different sets of potentially correlated predictors, thus ERA is well suited for the analysis of NSDUH data.

The literature on substance use, however, suggests that certain groups of US residents are dissimilar to others with respect to their socio-demographic characteristics, e.g., age, gender, ethnicity, etc. 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; Kandel et al., 2004; Robinson et al., 2006). Such patterns of heterogeneity depend on the specified statistical model, and moreover, it is difficult to know *a priori* which socio-demographic covariates to include. Unfortunately, ERA has no mechanism to account for such heterogeneity efficiently, thus being unable to examine whether the patterns of the relationships between variables vary across subgroups of observations differentiated by additional covariates.

In addition, ERA needs a more flexible modeling framework when the assumption of independent observations is violated. For example, the vast majority of substance users in the US population use more than one substance either concurrently or sequentially: the NSDUH

data show a positive association between cigarette and alcohol use, as well as a correlation between degree of alcohol use and rate of marijuana use among US residents. Considering the interdependence in the use of multiple substances, how to simultaneously analyze multiple correlated (or clustered) responses is a critical research question to address in ERA because the violation of independent observation assumption can invalidate any statistical inferences.

Thus, in this present research, two novel extensions of ERA are presented focusing on the theoretical underpinnings and empirical usefulness of the proposed extensions. The first extension, presented in Chapter 2, explores how to automatically identify heterogenous subgroups of respondents given a specified ERA model. To achieve this, ERA is combined with model-based recursive partitioning (MOB; Zeileis, Hothorn, & Hornik, 2008). A simulation study was conducted to evaluate the performance of the proposed method. From the application of the 2012 NSDUH data concerning nicotine dependence among US adults, the method could identify heterogeneous subgroups successfully based on a combination of sociodemographic covariates. This chapter was presented at the annual conference of the International Meeting of the Psychometric Society in July 2017, which was held in Zurich, Switzerland, and submitted to *British Journal of Mathematical and Statistical Psychology* in October 2019 and is currently the first round of revision.

The second extension, presented in Chapter 3, discusses how ERA incorporates generalized estimating equations (GEE; Liang & Zeger, 1986) to examine the effects of sets of predictors on multiple (and possibly correlated) responses simultaneously while relaxing the assumptions of correct specification of the covariance structure of the responses. This method also focusses on how to make the best use of ridge-type regularization to address any potential overfitting when many predictors per component are considered or when many

components influence the response variables. This new ERA extension was successfully applied to the analysis of rare genetic variants that are associated with multiple metabolic syndrome measures (Lee et al., 2019), as well as to the study of co-occurring recreational substance use among US adults (Kim et al., in press). Chapter 3 presents the major findings obtained by applying the method to the 2012 NSDUH data. This study was presented at the annual conference of the International Meeting of the Psychometric Society in July 2019, which was held in Santiago, Chile, and was accepted for publication in *Quantitative Psychology, IMPS 2019* (Kim et al., in press). The methodological details of this second ERA extension appear in Appendix C. Parameter Estimation in GEE-ERA, which was originally proposed in Lee et al. (2019). My contribution to this paper includes the development of the main algorithm, its implementation in a statistical package, and writing of the text.

Additionally, Chapter 4 introduces several new model evaluation metrics for ERA based on resampling methods, each of which aims to assess the predictive performance of ERA models on so-called out-of-sample data that are not used for parameter estimation. Although considerable work has been done in statistics and machine learning on the use of various cross-validation (CV) and bootstrap methods for out-of-sample prediction, to date, no research has applied these general tools to the ERA setting. Thus, I carried out simulation studies to evaluate the relative performance of different out-of-sample prediction error estimators for ERA. Simulation results, graphically examined using boxplots and error bars, illustrate which error estimator is the best to find the true model when mis-specified (i.e., overfitted) models are considered.

The final chapter summarizes the preceding chapters, highlighting the implications of new ERA extensions and discusses potential topics for future research.

References

Alin, A. (2009). Comparison of PLS algorithms when number of objects is much larger than number of variables. *Statistical Papers, 50*(4), 711–720. https://doi.org/10.1007/s00362-009-0251-7

Bentler, P. M., & Chou, C.-P. (1987). Practical Issues in Structural Modeling. *Sociological Methods & Research, 16*(1), 78–117. https://doi.org/10.1177/0049124187016001004

Bollen, K. A. (1989). *Structural Equations with Latent Variables*. Hoboken, NJ, USA: John Wiley & Sons, Inc. https://doi.org/10.1002/9781118619179

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

Bollen, K. A., & Davis, W. R. (2009). Causal Indicator Models: Identification, Estimation, and Testing. *Structural Equation Modeling: A Multidisciplinary Journal, 16*(3), 498–522. https://doi.org/10.1080/10705510903008253

Cheung, M. W. L., & Jak, S. (2016). Analyzing big data in psychology: A split/analyze/meta-analyze approach. *Frontiers in Psychology, 7*, 1–13. https://doi.org/10.3389/fpsyg.2016.00738

Choi, Kyung, M., Hwang, H., & Park, J. H. (2020). Bayesian Extended Redundancy Analysis: A Bayesian Approach to Component-based Regression with Dimension Reduction. *Multivariate Behavioral Research, 55*(1), 30–48. https://doi.org/10.1080/00273171.2019.1598837

Day, F. R., Ruth, K. S., Thompson, D. J., Lunetta, K. L., Pervjakova, N., Chasman, D. I., … Murray, A. (2015). Large-scale genomic analyses link reproductive aging to

hypothalamic signaling, breast cancer susceptibility and BRCA1-mediated DNA repair. *Nature Genetics*, *47*(11), 1294–1303. https://doi.org/10.1038/ng.3412

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

de Jong, S. (1993). SIMPLS: An alternative approach to partial least squares regression. *Chemometrics and Intelligent Laboratory Systems*, *18*(3), 251–263. https://doi.org/10.1016/0169-7439(93)85002-X

de Leeuw, J., Young, F. W., & Takane, Y. (1976). Additive structure in qualitative data: An alternating least squares method with optimal scaling features. *Psychometrika*, *41*(4), 471–503. https://doi.org/10.1007/BF02296971

De Mol, C., Giannone, D., & Reichlin, L. (2008). Forecasting using a large number of predictors: Is Bayesian shrinkage a valid alternative to principal components? *Journal of Econometrics*, *146*, 318–328. https://doi.org/10.1016/j.jeconom.2008.08.011

Deng, L., Yang, M., & Marcoulides, K. M. (2018). Structural Equation Modeling With Many Variables: A Systematic Review of Issues and Developments. *Frontiers in Psychology*, *9*, 1–14. https://doi.org/10.3389/fpsyg.2018.00580

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

Edwards, J. R., & Bagozzi, R. P. (2000). On the nature and direction of relationships between constructs and measures. *Psychological Methods*, *5*(2), 155–174. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/10937327

Efron, B., & Tibshirani, R. (1986). Bootstrap Methods for Standard Errors, Confidence Intervals, and Other Measures of Statistical Accuracy. *Statistical Science, 1*(1), 54–75. https://doi.org/10.1214/ss/1177013815

Enki, D. G., Trendafilov, N. T., & Jolliffe, I. T. (2013). A clustering approach to interpretable principal components. *Journal of Applied Statistics, 40*(3), 583–599. https://doi.org/10.1080/02664763.2012.749846

Farrar, D. E., & Glauber, R. R. (1967). Multicollinearity in Regression Analysis: The Problem Revisited. *The Review of Economics and Statistics, 49*(1), 92. https://doi.org/10.2307/1937887

Gauchi, J. P., & Chagnon, P. (2001). Comparison of selection methods of explanatory variables in PLS regression with application to manufacturing process data. In *Chemometrics and Intelligent Laboratory Systems* (Vol. 58, pp. 171–193). Elsevier. https://doi.org/10.1016/S0169-7439(01)00158-7

Geladi, P., & Kowalski, B. R. (1986). Partial least-squares regression: a tutorial. *Analytica Chimica Acta, 185*(C), 1–17. https://doi.org/10.1016/0003-2670(86)80028-9

Gusev, A., Ko, A., Shi, H., Bhatia, G., Chung, W., Penninx, B. W. J. H., … Pasaniuc, B. (2016). Integrative approaches for large-scale transcriptome-wide association studies. Nature Genetics, 48(3), 245–252. https://doi.org/10.1038/ng.3506

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

Hotelling, H. (1957). The Relations of the Newer Multivariate Statistical Methods to Factor Analysis. *British Journal of Statistical Psychology, 10*(2), 69–79. https://doi.org/10.1111/j.2044-8317.1957.tb00179.x

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.-H., & Lim, J. (2015a). Generalized Functional Extended Redundancy Analysis. *Psychometrika, 80*(1), 101–125. https://doi.org/10.1007/s11336-013-9373-x

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

Hwang, H., & Takane, Y. (2014). *Generalized structured component analysis: A component-based approach to structural equation modeling* (1st ed.). New York: Chapman and Hall/CRC. https://doi.org/10.1201/b17872

Jackson, D. L. (2001). Sample Size and Number of Parameter Estimates in Maximum Likelihood Confirmatory Factor Analysis: A Monte Carlo Investigation. *Structural Equation Modeling: A Multidisciplinary Journal, 8*(2), 205–223. https://doi.org/10.1207/S15328007SEM0802_3

Jolliffe, I. T. (1982). A Note on the Use of Principal Components in Regression. *Journal of the Royal Statistical Society. Series C (Applied Statistics), 31*(3), 300–303. https://doi.org/10.2307/2348005

Jöreskog, K. G. (1970). A general method for analysis of covariance structures. *Biometrika, 57*(2), 239–251. https://doi.org/10.1093/biomet/57.2.239

Jöreskog, K. G. (1973). A generating method for estimating a linear structural equation system. In A. S. Goldberger & O. D. Duncan (Eds.), *Structural Equation Models in the*

Social Sciences (pp. 85–112). Seminar Press. https://doi.org/10.1002/j.2333-8504.1970.tb00783.x

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

Kendall, M. G. (1957). *A Course in Multivariate Analysis*. London: Charles Griffen & Co. Retrieved from https://www.amazon.com/Course-Multivariate-Analysis-M-Kandall/dp/B001PKBO90

Kozberg, M. G., Chen, B. R., DeLeo, S. E., Bouchard, M. B., & Hillman, E. M. C. (2013). Resolving the transition from negative to positive blood oxygen level-dependent responses in the developing brain. *Proceedings of the National Academy of Sciences of the United States of America, 110*(11), 4380–4385. https://doi.org/10.1073/pnas.1212785110

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, 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

Liang, K. Y., & Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika, 73*(1), 13–22. https://doi.org/10.1093/biomet/73.1.13

MacCallum, R. C., & Browne, M. W. (1993). The use of causal indicators in covariance structure models: some practical issues. *Psychological Bulletin*, *114*(3), 533–541. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/8272469

Maitra, S., & Models, J. Y. (2008). Principle component analysis and partial least squares: Two dimension reduction techniques for regression. *Applying Multivariate Statistical Models*, *79*, 79–90. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.473.4340&rep=rep1&type=pdf#page=81

McDonald, R. P. (1996). Path analysis with composite Variables. *Multivariate Behavioral Research*, *31*(2), 239–270. https://doi.org/10.1207/s15327906mbr3102_5

Mehmood, T., & Ahmed, B. (2016). The diversity in the applications of partial least squares: an overview. *Journal of Chemometrics*, *30*(1), 4–17. https://doi.org/10.1002/cem.2762

Moustafa, A. A., Diallo, T. M. O., Amoroso, N., Zaki, N., Hassan, M., & Alashwal, H. (2018). Applying Big Data Methods to Understanding Human Behavior and Health. *Frontiers in Computational Neuroscience*, *12*(84), 1–4. https://doi.org/10.3389/fncom.2018.00084

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

Shmueli, G. (2010). To Explain or to Predict? *Statistical Science*, *25*(3), 289–310. https://doi.org/10.1214/10-STS330

Smith, G., & Campbell, F. (1980). A critique of some ridge regression methods. *Journal of the American Statistical Association, 75*(369), 74–81. https://doi.org/10.1080/01621459.1980.10477428

Smith, & Sasaki. (1979). Decreasing multicollinearity: A method for models with multiplicative functions. *Sociological Methods & Research, 8*(1), 35–56. https://doi.org/10.1177/004912417900800102

Spano, V. R., Mandell, D. M., Poublanc, J., Sam, K., Battisti-Charbonney, A., Pucci, O., … Mikulis, D. J. (2013). CO2 Blood Oxygen Level–dependent MR Mapping of Cerebrovascular Reserve in a Clinical Population: Safety, Tolerability, and Technical Feasibility. *Radiology, 266*(2), 592–598. https://doi.org/10.1148/radiol.12112795

Stone, M. (1974). Cross-Validatory Choice and Assessment of Statistical Predictions. *Journal of the Royal Statistical Society, 36*(2), 111–147. https://doi.org/10.2307/2984809

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

Tian, T. S., Wilcox, R. R., & James, G. M. (2010). Data reduction in classification: A simulated annealing based projection method. *Statistical Analysis and Data Mining: The ASA Data Science Journal, 3*(5), 319–331. https://doi.org/10.1002/sam.10087

United States Department of Health and Human Services. Substance Abuse and Mental Health Services Administration. Center for Behavioral Health Statistics and Quality. (2013). National Survey on Drug Use and Health Database. Inter-university Consortium

for Political and Social Research (ICPSR) [distributor]. https://doi.org/10.3886/ICPSR35509.v1

Wold, H. (1966). Estimation of principal components and related models by iterative least squares. In P. R. Krishnajah (Ed.), *Multivariate analysis* (pp. 391-420). New York: Academic Press.

Chapter 2. Model-Based Recursive Partitioning of Extended Redundancy Analysis

Manuscript: Kim, S. & Hwang, H. Model-Based Recursive Partitioning of Extended Redundancy Analysis with an Application to Nicotine Dependence among US adults. Submitted to *British Journal of Mathematical and Statistical Psychology* (2019).

Abstract

Extended redundancy analysis (ERA) is used to reduce multiple sets of predictor variables 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 effects of predictor components on a response variable vary across subgroups differentiated by covariates. To address this issue, we propose to combine ERA with model-based recursive partitioning in a single framework. This method aims to partition observations into heterogeneous subgroups recursively based on a set of covariates and to apply ERA to each subgroup simultaneously. It can show how the impacts of components on a response variable differ across covariate-dependent subgroups. Moreover, it produces a tree diagram that aids in visualizing a hierarchy of 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 covariates, each of which yielded different effects of components on regression coefficients.

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

2.1. 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 et al., 2015), and multivariate ERA for the analysis of multiple correlated responses (Lee et al., 2019, 2018).

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 et al., 2006a; Kandel et al., 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, 2015a). 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 socioeconomic status (SES) groups (Brandmaier, von Oertzen, McArdle, & Lindenberger, 2013; McArdle & Epstein, 1987; Von Stumm & Plomin, 2015). Although such covariate-dependent heterogeneity or observations is prevalent in practice, ERA has no mechanism to account for this heterogeneity efficiently, thus being unable to examine whether the relationships between components and a response variable vary across subgroups of observations differentiated by additional covariates.

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 investigate whether the effects of components on a response variable are different across covariate-dependent subgroups. 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, 2016a; Strobl et al., 2015a; Strobl, Wickelmaier, & Zeileis, 2011).

The proposed method, called MOB-ERA hereinafter, begins by fitting an ERA model to entire observations, producing a single set of the ERA parameter estimates, and then successively inspects whether there are substantial changes in the effects of components on the response variable across covariate-dependent subgroups. This is achieved through the so-called parameter instability test in MOB that uses the score for the log-likelihood function of

the ERA model, as will be discussed in detail in the Methods section. The proposed 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 to review 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 the proposed method. We then 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 that the method may identify heterogeneous subgroups successfully based on a combination of sociodemographic covariates. We finally discuss the implications of the method and potential topics for future research.

2.2. Methods

2.2.1. ***Parametric ERA***

Assume that there are K different sets of predictors, each of which consists of P_k predictors ($k = 1, \cdots, K$). Let x_{ikp} denote the ith value of the pth variable in the kth predictor set ($i = 1, \cdots, N$; $p = 1, \cdots, P_k$) 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 *i*th component score of the *k*th component, which is the sum of weighted predictors for the *i*th observation in the *k*th predictor set. Let b_k denote the regression coefficient relating the *k*th component to the response variable. Let η_i and $g(\cdot)$ denote the *i*th 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

$$\begin{aligned} g(\mu_i) &= \eta_i \\ &= \sum_{k=1}^{K} \left[\sum_{p=1}^{P_k} x_{ikp} w_{kp} \right] b_k = \boldsymbol{x}_i' \mathbf{W} \boldsymbol{b} \\ &= \sum_{k=1}^{K} f_{ik} b_k = \boldsymbol{f}_i \boldsymbol{b}, \end{aligned} \tag{2.2.1}$$

where $\mathbf{W} = \text{diag}(\boldsymbol{w}_1, \cdots, \boldsymbol{w}_K)$, $\boldsymbol{f}_i' = (f_{i1}, \cdots, f_{iK})$, and $\boldsymbol{b} = (b_1, \cdots, b_K)'$. As shown in (2.2.1), 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 2.2.1 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, the **W** and ***b*** are given as

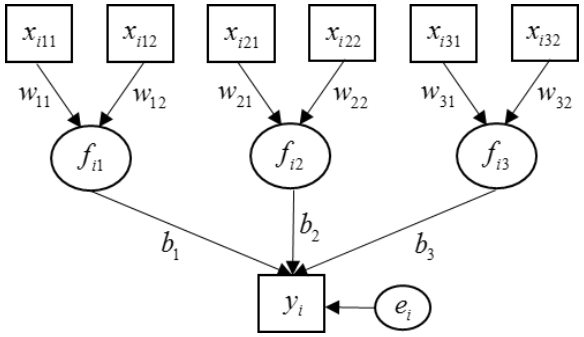


Figure 2.2.1. An exemplary parametric ERA model

$$\mathbf{W} = \text{diag}(\boldsymbol{w}_1, \boldsymbol{w}_2, \boldsymbol{w}_3) = \begin{pmatrix} w_{11} & 0 & 0 \\ w_{12} & 0 & 0 \\ 0 & w_{21} & 0 \\ 0 & w_{22} & 0 \\ 0 & 0 & w_{31} \\ 0 & 0 & w_{32} \end{pmatrix} \quad \text{and} \quad \boldsymbol{b} = \begin{pmatrix} b_1 \\ b_2 \\ b_3 \end{pmatrix}.$$

In the present paper, we assume that y_i is independently distributed with the mean μ_i and has the probability density (or mass) function of the form

$$f(y_i; \theta_i, \phi) = \exp\left(\frac{y_i \theta_i - \beta(\theta_i)}{\alpha(\phi)} + \gamma(y_i, \phi) \right),$$

for known functions $\alpha(\cdot)$, $\beta(\cdot)$, and $\gamma(\cdot)$, where θ_i is the natural (or canonical) parameter that can be expressed as some function of μ_i and ϕ is a constant dispersion parameter. If the dispersion parameter ϕ is known, the above equation belongs to the exponential family with the canonical parameter θ_i (McCullagh & Nelder, 1989, Chapter 2; Nelder & Wedderburn, 1972). Many commonly used distributions, such as the normal, gamma, binomial, and Poisson, are in this family. The log-likelihood function for N observations from the exponential family is generally written as a function of $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_N)'$, i.e.,

$$\ell(\boldsymbol{\theta}; y_1, \ldots, y_N) = \sum_{i=1}^{N} \log f(y_i; \theta_i, \phi) = \sum_{i=1}^{N} \frac{y_i \theta_i - \beta(\theta_i)}{\alpha(\phi)} + \sum_{i=1}^{N} \gamma(y_i, \phi)$$

in its "natural" form, where the function is parametrized by its natural (or canonical) parameters. We seek to maximize the $\ell(\boldsymbol{\theta}; y_1, \ldots, y_N)$ over $\boldsymbol{\theta}$, thus the log-likelihood can be written as

$$\ell(\boldsymbol{\theta}; y_1, \ldots, y_N) = \sum_{i=1}^{N} y_i \theta_i - \beta(\theta_i),$$

where the terms that do not depend on $\boldsymbol{\theta}$ are discarded. Suppose we consider 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

$$\ell(\boldsymbol{\theta}_{\text{ERA}}; y_1, \ldots, y_N) = \sum_{i=1}^{N} y_i \boldsymbol{x}_i' \mathbf{W} \boldsymbol{b} - \beta(\boldsymbol{x}_i' \mathbf{W} \boldsymbol{b}) = \sum_{i=1}^{N} y_i \boldsymbol{f}_i \boldsymbol{b} - \beta(\mathbf{f}_i \boldsymbol{b}), \qquad (2.2.2)$$

where $\boldsymbol{\theta}_{\text{ERA}}$ denotes a (P+K) by 1 vector that stacks $\boldsymbol{w}_k$'s 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 (2.2.2) via IRLS is equivalent to minimizing the following generalized least-squares criterion (Hwang et al., 2015; Lee et al., 2016)

$$\varphi_{(w_{kp}, b_k)} = \sum_{i=1}^{N} \omega_i (z_i - \sum_{k=1}^{K} \left[\sum\nolimits_{p=1}^{P_k} x_{ikp} w_{kp} \right] b_k)^2 = \sum_{i=1}^{N} \omega_i (z_i - \sum_{k=1}^{K} f_{ik} b_k)^2, \qquad (2.2.3)$$

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 (2.2.3) (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 Appendix B. Estimation and Inference in Parametric ERA for a detailed description of the algorithm.

2.2.2. ***Recursive Partitioning of ERA***

As discussed earlier, ERA is currently unable to capture covariate-dependent heterogeneity, ignoring potential subgroup-specific relationships between components and a response variable based on additional covariates. As in other MOB extensions, 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, Zeileis, & Hothorn, 2016b; Thomas, Bornkamp, & Seibold, 2018).

To identify heterogeneous subgroups based on a given set of covariates in ERA, we propose MOB-ERA that combines ERA with MOB in a unified manner. More specifically, the so-called parameter instability test in MOB (Seibold et al., 2016a; 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. This test focuses on whether there are statistically significant changes or instabilities in parameter estimates across subgroups derived from a partitioning covariate.

Let ψ_i denote the score contribution of the ith observation or the gradient of the log-likelihood contribution of the ith observation with respect to the model parameters, i.e.,

$$\psi_i = s\left(\boldsymbol{\theta}_{\text{ERA}}; (y, \boldsymbol{x})_i\right) = \frac{\partial \ell(\boldsymbol{\theta}_{\text{ERA}}; (y, \boldsymbol{x})_i)}{\partial \boldsymbol{\theta}_{\text{ERA}}}. \tag{2.2.4}$$

Then, the empirical score contribution of the ith individual is

$$\hat{\psi}_i = s\left(\hat{\boldsymbol{\theta}}_{\text{ERA}}; (y, \boldsymbol{x})_i\right) = \left.\frac{\partial \ell(\boldsymbol{\theta}_{\text{ERA}}; (y, \boldsymbol{x})_i)}{\partial \boldsymbol{\theta}_{\text{ERA}}}\right|_{\hat{\boldsymbol{\theta}}_{\text{ERA}}}, \tag{2.2.5}$$

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 $\hat{\psi}_i$'s 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 empirical scores 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.

The procedures using the score contributions in (2.2.5) assume that both sets of the ERA parameters, i.e., component weights and regression coefficients, are to vary across subgroups. This is technically possible and might be of interest depending on research questions (e.g., does a predictor contribute differently to forming its component in two gender groups?). However, in the present paper, we estimate component weights once based on the entire observations and then consider them fixed for all subsequent subgroups, while estimating regression coefficients freely across the subgroups. This assures that components or regressors in an ERA model convey the same meanings across different subgroups of the resulting tree, so that the regression coefficients in one subgroup can be compared with those in another. That is, we first fit a specified ERA model to all observations and obtain the global component weight estimates $\hat{\mathbf{W}}$ and component scores $\boldsymbol{f}_i = \boldsymbol{x}_i\hat{\mathbf{W}}$. Then, we use the following score contributions, rather than using (2.2.5), for the data partitioning procedures described above.

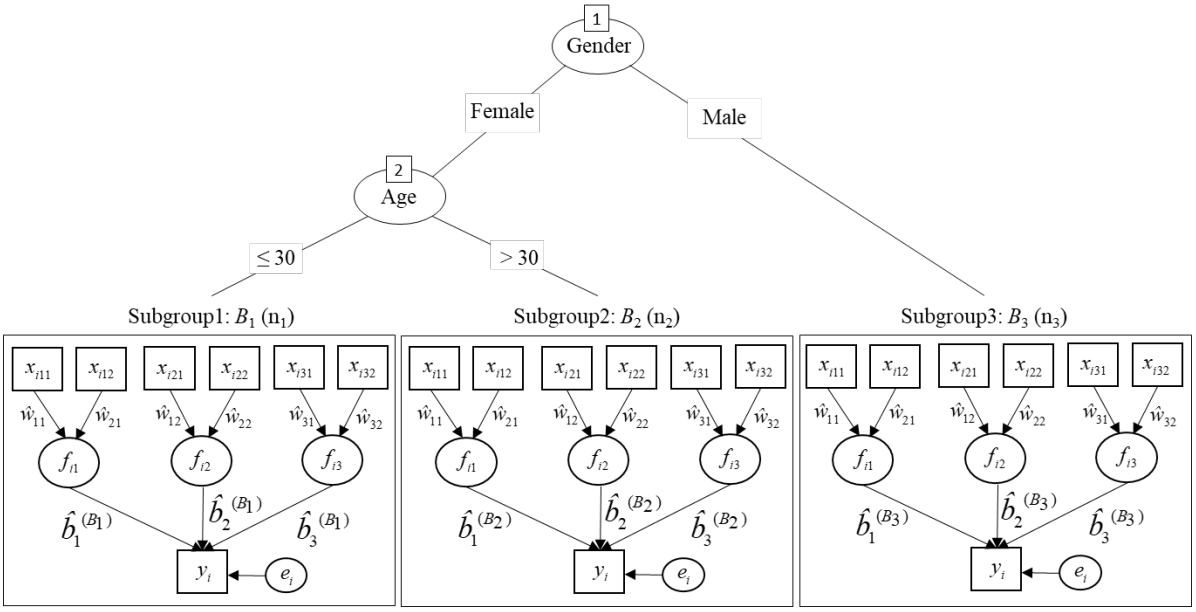


Figure 2.2.2. An illustrative example of MOB-ERA

$$\hat{\psi}_i = s\left(\hat{\boldsymbol{b}}; (y, \boldsymbol{f})_i\right) = \left.\frac{\partial \ell(\boldsymbol{b}; (y, \boldsymbol{f})_i)}{\partial \boldsymbol{b}}\right|_{\hat{\boldsymbol{b}}}. \quad (2.2.6)$$

Figure 2.2.2 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 2.2.1, entire 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.

2.2.3. ***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., 2016a; 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:

$$\text{AIC}^{(\text{Parent node})} = -2 \cdot \ell(\hat{\boldsymbol{\theta}}_{\text{ERA}}^{(\text{Parent node})}) + 2 \cdot h^{(\text{Parent node})} \tag{2.2.7}$$

and

$$\text{AIC}^{(\text{Subsequent nodes: A and B})} = -2 \cdot (\, \ell(\hat{\boldsymbol{\theta}}_{\text{ERA}}^{(\text{Node A})}) + \ell(\hat{\boldsymbol{\theta}}_{\text{ERA}}^{(\text{Node B})}) \,) + 2 \cdot (h^{(\text{Node A})} + h^{(\text{Node B})}), \tag{2.2.8}$$

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 (2.2.7) represents the relative amount of information assuming a single set of parameter estimates (*simpler model of homogeneity*), where the AIC in (2.2.8) quantifies the information assuming different sets of parameter estimates (*complex model of heterogeneity*). For example, in Figure 2.2.2, 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.

2.3. A Simulation Study

We investigated a Type I error rate, power, and classification accuracy of the proposed method. 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 under the proposed method. We also investigated how well and accurately the proposed method could detect parameter instability, thereby identifying the subgroups derived from pre-specified partitioning covariates correctly.

2.3.1. ***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.

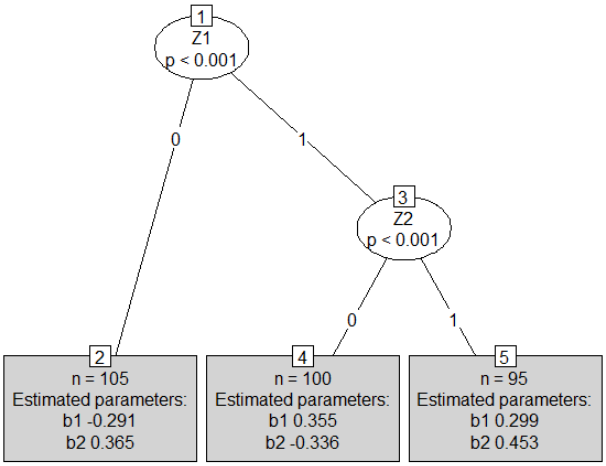


Figure 2.3.1. An example of simulated data when $N = 300$ and $n_1 = n_2 = n_3$

We considered six different sample sizes: $N = 90$, 120, 180, 300, 600, and 900. As depicted in Figure 2.3.1, 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 binary partitioning covariates, Z_1 and Z_2. Both covariates were randomly sampled from a binomial distribution, $Z_1 \sim B(N, p_1+(1-p_1)\cdot\delta)$ and $Z_2 \sim B(N, p_2+(1-p_2)\cdot\delta)$, where $p_1 = P(Z_1 = 0) = n_1/N$, $p_2 = P(Z_2 = 0|Z_1 = 1) = n_2/(n_2+n_3)$, and δ is the instability control parameter taking the value of either 1 or 0. In this study, we used δ to generate either a homogeneity case for evaluating the Type I error ($\delta = 1$) or a heterogeneity case for evaluating power and accuracy ($\delta = 0$), i.e.,

$$\begin{cases} Z_1 \sim B(N, p_1),\ Z_2 \sim B(N, p_2) & \text{if } \delta = 0 \\ Z_1 \sim B(N, 1),\ Z_2 \sim B(N, 1) & \text{if } \delta = 1. \end{cases}$$

Note that when $\delta = 1$, both success probabilities of Z_1 and Z_2 are equal to 1, thus generating one homogenous subgroup. We also included a noise covariate Z_3 that was completely unrelated to the subgroups to examine whether the proposed method could accurately select the correct covariate when partitioning data. The noise covariate Z_3 was sampled from a uniform distribution between 0 and 1.

Under the heterogeneity case, the subgroups differed by the partitioning covariates had different signs (directions) of the two regression coefficients as follows:

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

Note that the difference in the magnitude of regression coefficient b_2 was affected by the value of R^2. Finally, we varied the number of observations for each subgroup as follows:

$$(n_1, n_2, n_3) = \begin{cases} (1/3,\ 1/3,\ 1/3) \cdot N, \text{ Balanced} \\ (1/2,\ 1/4,\ 1/4) \cdot N, \text{ Moderately unbalanced} \\ (2/3,\ 1/6,\ 1/6) \cdot N, \text{ Considerably unbalanced.} \end{cases}$$

As shown above, in the balanced condition, the number of observations for each subgroup was all equal, whereas in the unbalanced conditions, one group size was larger than the others.

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 1,000 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, moderately-, or considerably-unbalanced). We applied the proposed method 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.

Table 2.3.1. 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	# obs. for each subgroup	R^2	90	120	180	300	600	900
Type I error	(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	.2	.91	.91	.93	1.00	1.00	1.00
		.4	.93	.96	1.00	1.00	1.00	1.00
		.6	.94	.98	1.00	1.00	1.00	1.00
	Moderately unbalanced	.2	.79	.84	.91	1.00	1.00	1.00
		.4	.80	.85	.99	1.00	1.00	1.00
		.6	.80	.85	1.00	1.00	1.00	1.00
	Considerably unbalanced	.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
Cramer's V	Balanced	.2	.90	.90	.92	.97	.97	.99
		.4	.91	.91	.95	1.00	.98	.99
		.6	.92	.92	.99	.99	.99	.99
	Moderately unbalanced	.2	.80	.85	.90	.95	.94	.89
		.4	.87	.81	.93	.95	.95	.90
		.6	.86	.88	.95	.99	.99	.99
	Considerably unbalanced	.2	.81	.83	.84	.91	.93	.88
		.4	.88	.87	.86	.96	.95	.89
		.6	.85	.86	.95	.99	.99	.99

2.3.2. ***Results***

In this study, an empirical Type I error was calculated by counting how many of the samples were falsely partitioned under the homogeneity case ($\delta = 1$). Table 2.3.1 presents the empirical Type I error rates across the different sample sizes and the different values of R^2. In all the conditions, the proposed method 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, the proposed method 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.

Table 2.3.1 also provides the empirical power of the proposed method over different sample sizes and R^2 values under the heterogeneity case ($\delta = 0$), 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 1,000 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, the proposed method was able to detect instabilities beyond a power threshold of .9 across all the sample sizes and R^2 values. Under the moderately and considerably unbalanced conditions, 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., $R^2 = .6$). To ensure an adequate level of power of the proposed method in small samples, therefore, the size of any subgroup should not be too dominant. Although not reported in Table 2.3.1, 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 2.3.1 also displays the average Cramér's V values for the different sample sizes and R^2 values under the heterogeneity case ($\delta = 0$). Under the balanced condition, on average, the Cramér's V increased with the sample size and R^2. Moreover, Cramér's V were all around .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 increased ($N > 180$) and/or R^2 became large. Interestingly, Cramér's V decreased again when $N = 900$ because the proposed method ended up partitioning data into more than the pre-specified number of subgroups. This suggests that pruning might be necessary in large samples to avoid such overfitting.

2.4. An Empirical Application

We applied the proposed method 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 2.4.1 presents a description of all the variables and their summary statistics. It also shows which component is associated with which predictors. Figure 2.4.1 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.

Table 2.4.1. A description of variables and summary statistics for the 2012 NSDUH data

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).

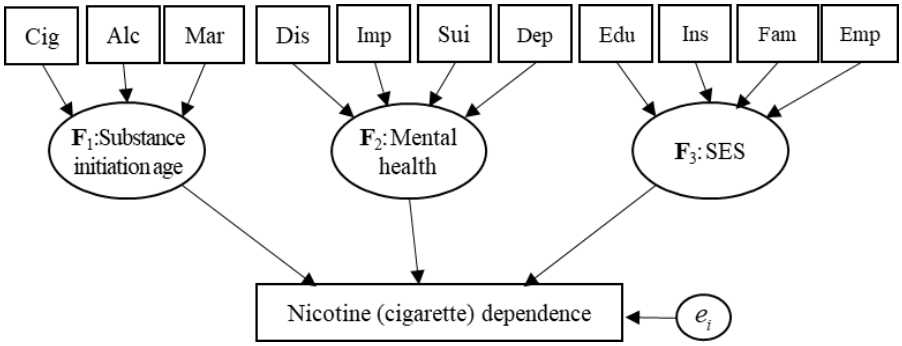


Figure 2.4.1. The ERA model for the 2012 NSDUH data (Variable names are consistent with those in Table 2.4.1)

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 et al., 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.

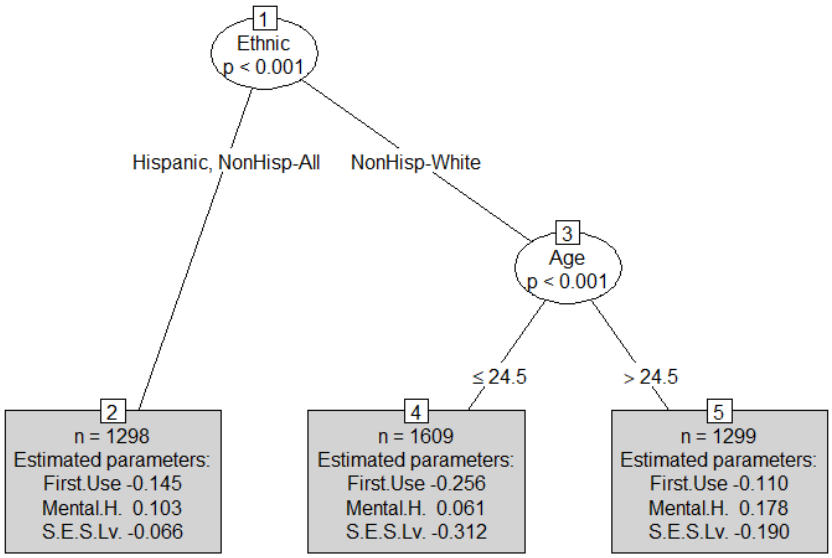
Table 2.4.2 presents the estimated component weights, their standard errors, and *p*-values for all the respondents. The first three columns of the table show the results obtained from the training set. As shown in the table, the component weight estimate for age of first cigarette use (w_{11}) was positive and statistically significant, indicating that cigarette initiation contributed to forming F_1, substance initiation age, in explaining the degree of nicotine dependence. Neither alcohol nor marijuana initiation age was statistically significant. The estimate for the level of functional impairment in daily life (w_{22}) was positively and statistically significantly related to F_2, mental health status, whereas the rest of the predictors for this component set was not. In the last predictor set, the weight estimates for three predictors, including education level (w_{31}), insurance (w_{32}), and job status (w_{34}), were positive

Table 2.4.2. The component weight estimates (Est.), and their standard errors (S.E.) and *p*-values from MOB-ERA for the 2012 NSDUH data.

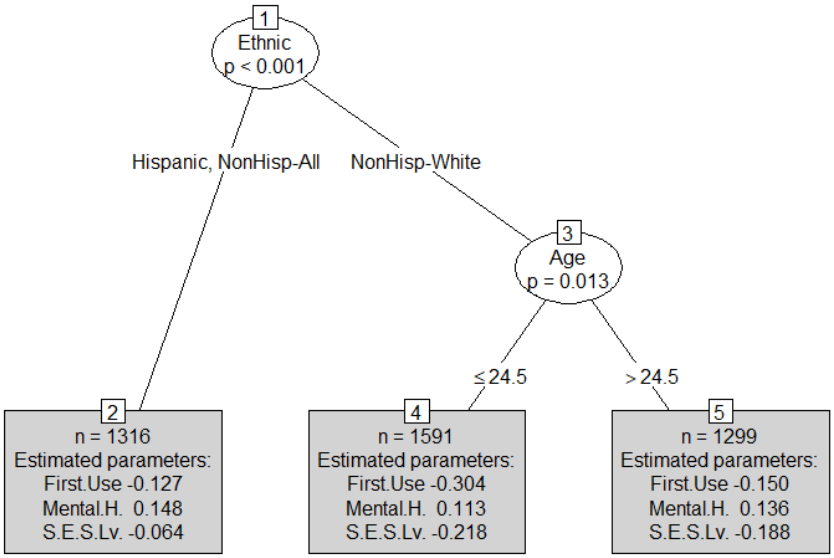
		(a) Training set			(b) Test set		
Components	Predictors	Est.	S.E.	*p*-val	Est.	S.E.	*p*-val
$\mathbf{F}_1$	Cigarette initiation (w_{11})	.95	.11	.00	1.02	.10	.00
	Alcohol initiation (w_{12})	.18	.11	.12	-.25	.10	.12
	Marijuana initiation (w_{13})	-.08	.11	.45	.12	.10	.24
$\mathbf{F}_2$	Distress level (w_{21})	.30	.19	.14	.46	.15	.01
	Impairment (w_{22})	.64	.18	.00	.60	.14	.00
	Suicidal thought (w_{23})	.06	.15	.72	.00	.13	.98
	Depression (w_{23})	.18	.17	.30	.09	.13	.53
$\mathbf{F}_3$	Education (w_{31})	.77	.08	.00	.74	.10	.00
	Insurance (w_{32})	.36	.08	.00	.36	.09	.00
	Family income (w_{33})	.04	.08	.66	.05	.10	.63
	Employment Status (w_{34})	.29	.08	.00	.30	.10	.01

and statistically significant, contributing to determining F_3, SES. The family income was not statistically significantly related to F_3. As shown in the last three columns of the table, similar results were obtained from the test set.

Given the component weight estimates, the proposed method identified potentially heterogeneous subgroups, which might exhibit distinct effects of the three components on the degree of nicotine dependence. As partitioning covariates, we considered four sociodemographic variables: age, gender, marital status, and ethnicity. Refer to Table 2.4.1 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



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



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

Figure 2.4.2. 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.

when there are continuous covariates, categorical covariates with multiple levels, and/or a number of covariates at the same time (Strobl, Kopf, & Zeileis, 2015b; Su, Tsai, Wang, Nickerson, & Li, 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 $\alpha = .05$) 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 2.4.2 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.

Table 2.4.3 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 24.5 and over 24.5) 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 2.4.2), the final hierarchy of the partitioning covariates was the same

Table 2.4.3. A summary of the parameter instability tests for the 2012 NSDUH data

		Age		Gender		Marital Status		Ethnicity	
	Node	Statistic	*p*-value	Statistic	*p*-value	Statistic	*p*-value	Statistic	*p*-value
(a) Training set	1	36.38	.00	7.89	.18	21.57	.04	43.14	.00
	3	42.53	.00	.73	.99	30.15	.00	0 [a]	-
(b) Test set	1	20.01	.01	4.98	.53	14.41	.37	36.14	.00
	3	17.89	.01	3.02	.77	18.69	.08	0 [a]	-

[a] Node 3 is ethnically homogeneous.

Table 2.4.4. The regression coefficient estimates (Est.), and their standard errors (S.E.) and *p*-values from MOB-ERA for the 2012 NSDUH data

		F$_1$: Substance initiation			**F**$_2$: Mental health status			**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)	-.15	.03	.00	.10	.03	.00	-.07	.03	.01
	4 (N=1,609)	-.26	.03	.00	.06	.02	.01	-.31	.03	.00
	5 (N=1,299)	-.11	.02	.00	.18	.03	.00	-.19	.03	.00
(b) Test set	2 (N=1,316)	-.13	.03	.00	.15	.03	.00	-.06	.03	.01
	4 (N=1,591)	-.30	.03	.00	.11	.02	.00	-.22	.03	.00
	5 (*N*=1,299)	-.15	.02	.00	.14	.03	.00	-.19	.03	.00

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.

Table 2.4.4 shows the estimated regression coefficients and their standard errors per subgroup. The estimates are also displayed at each terminal node in Figure 2.4.2. 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 24.5 (Node 4), compared to the other groups. Moreover, SES had the smallest effect on the nicotine dependence in the non-White respondents (Node 2), whereas it had the largest effect in the group of older Whites aged over 24.5 (Node 5). This older Whites group also showed the largest effect of mental health status on nicotine dependence among the three groups. Again, similar findings were obtained from the test set.

2.5. 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 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. The simulation study showed that the proposed method seemed to control for the Type I error rate reasonably well over the whole range of regression coefficients 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 proposed method also showed better performance in empirical power and classification accuracy, particularly when the number of observations was equal for all subgroups.

We also demonstrated how the proposed method could identify covariate-dependent heterogeneous subgroups, using a well-known national survey dataset in the US. When partitioning the data based on a 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 automatically derived, 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 the proposed method was reliable in detecting heterogeneous subgroups.

The present study proposed a new extension of ERA and demonstrated its empirical utility using the 2012 NSDUH data. As with many other recursive partitioning methods, however, 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). In our empirical application, similar solutions were obtained from both training and validation sets. Nevertheless, it would be worthwhile to apply the proposed method to a wide range of real data to investigate such variability of solutions more carefully. Moreover, it may be necessary to technically refine the method to alleviate this potential 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.

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

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

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.-H., & Lim, J. (2015a). Generalized Functional Extended Redundancy Analysis. *Psychometrika, 80*(1), 101–125. https://doi.org/10.1007/s11336-013-9373-x

Hwang, H., Suk, H. W., Takane, Y., Lee, J., & Lim, J. (2015b). 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

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

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

Seibold, H., Zeileis, A., & Hothorn, T. (2016b). 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. (2015a). 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., Kopf, J., & Zeileis, A. (2015b). 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

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

Chapter 3. Regularized Extended Redundancy Analysis via Generalized Estimating Equations

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

Abstract

According to the National Survey on Drug Use and Health (NSDUH), the co-use of recreational substances is prevalent in the US population and engenders serious public health consequences. Additionally, substance use is an example of a complex social phenomenon that involves a large number of potentially correlated predictors. Considering the interdependence in the use of cigarettes, alcohol, and marijuana among US adults, the purpose of this study is to investigate simultaneously the effects of multiple sets of predictors (regarding substance initiation age, mental health status, and socioeconomic status) on the use of these three substances. For this, we applied a recently proposed extension of extended redundancy analysis (ERA), named GEE-ERA, to the 2012 NSDUH data. ERA performs data reduction and linear regression simultaneously, producing a simpler description of directional relationships between multiple sets of predictors and response variables. The new extension, GEE-ERA, combines ERA with generalized estimating equations (GEE) to enable fitting a regression on a set of correlated responses with unknown correlation structure. This method also adopts ridge-type regularization to address any potential overfitting, while the strength of the regularization is determined automatically through cross-validation. The major findings obtained by applying GEE-ERA to the 2012 NSDUH data are: (1) Earlier substance use was associated with greater current use of both cigarettes and alcohol; (2) worse mental health status influenced greater marijuana use, only; and (3) a lower level of SES was associated with higher levels of both cigarette and marijuana use.

Keywords: Co-occurring substance use, substance initiation age, mental health, socioeconomic status, component-based dimension reduction, extended redundancy analysis, generalized estimating equations, regularization

3.1. Background

The current substance use epidemic in the US leads to adverse public health consequences, such as drugged driving (National Institute on Drug Abuse, 2019) and smoking- or alcohol-related cancers (U.S. Department of Health and Human Services, 2004). According to the 2012 National Survey on Drug Use and Health (NSDUH), an estimated 62% of Americans aged 12 and older used at least one recreational psychoactive substance (i.e., tobacco, alcohol, or illicit drug) within the past year, including 9% who met the criteria for substance abuse disorder (United States Department of Health and Human Services. Substance Abuse and Mental Health Services Administration. Center for Behavioral Health Statistics and Quality, 2013). Moreover, the same 2012 NSDUH data show a positive association between cigarette and alcohol use, as well as a correlation between degree of alcohol use and rate of illicit drug use (of which marijuana use accounts for the vast majority) (United States Department of Health and Human Services. Substance Abuse and Mental Health Services Administration. Center for Behavioral Health Statistics and Quality, 2013). Considering that the vast majority of substance users (91% in 2012) use more than one substance, either concurrently or sequentially, a statistical model that simultaneously analyzes use of multiple substances would provide a more complete representation of the phenomenon of substance co-use among US adults.

Further complicating the study of substance use among US adults is the large number of predictors that have been demonstrated in previous studies to explain the use of one or more substances (e.g., Daza et al., 2006; Hu, Davies, & Kandel, 2006b; Kandel et al., 2004; Robinson et al., 2006). Categories of such predictors include: (1) substance initiation age (i.e., age of first cigarette, alcohol, and/or marijuana use), (2) indicators of mental health (e.g., major depressive episode during past year, daily functional impairment level, etc.), and

(3) indicators of socioeconomic status (SES; education level, health insurance coverage, family income, employment status). Considering such a high-dimensional set of predictors, the major difficulty in investigating the effect of numerous predictors on the concurrent use of substances is the lack of statistical methods capable of providing a comprehensible description of directional relationships among many sets of variables, without suffering from potential multicollinearity issues.

Thus, in the present work, we use regularized extended redundancy analysis (Takane & Hwang, 2005) combined with generalized estimating equations (Liang & Zeger, 1986) to investigate associations between the aforementioned predictor sets and correlated use of multiple substances. ERA is a statistical method that relates multiple sets of predictors to response variables. In ERA, a component is extracted from each set of predictor variables such that it accounts for the maximum variation of response variables. In this regard, ERA performs data reduction and linear regression simultaneously, producing a simpler description of directional relationships between multiple sets of predictors and response variables. Recently, a new extension of ERA was proposed for the analysis of clustered or correlated response variables (Lee et al., 2019). In this extension, GEE is combined with ERA to model response variables with an unknown correlation structure. This new method, called GEE-ERA hereinafter, can handle different types of response variables (e.g., continuous, binary, or count) that are assumed to follow an exponential family distribution. The method also incorporates ridge-type regularization to address potential overfitting when many predictors per component are considered or when many components influence the response variables. The regularization strength is determined automatically using cross-validation (CV).

The remainder of the paper is organized as follows. We begin by briefly reviewing GEE-ERA focusing especially on its advantages for the analysis of co-occurring substance use in the US. We then apply the method to data from the 2012 National Survey on Drug Use and Health (NSDUH), an annual survey that provides extensive statistical information on the use of recreational psychoactive substances and various associated sociopsychological variables. This application shows that GEE-ERA can identify meaningful predictors while taking into account the correlation structure of nicotine, alcohol, and marijuana use and preventing overfitting by the regularization strategy. We conclude by discussing the implications of the method and topics for future research.

3.2. Method

3.2.1. ***Model Specification***

In GEE-ERA (Lee et al., 2019), we assume that there are Q response variables and K different sets of predictors, each of which consists of P_k predictors (k =1, ⋯, K). Let y_{iq} denote the value of the qth response variable measured on the ith respondent (i = 1, ⋯, N; q = 1, ⋯, Q). We assume that y_{iq} follows an exponential family distribution with a mean μ_{iq} and variance $\phi\sigma_{iq}^2$, where ϕ is a dispersion parameter which may or may not be of substantive interest. Let w_{kp} denote the component weight assigned to x_{ikp}. 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 predictor variables for the ith observation in the kth predictor set. Let β_{kq} denote the regression coefficient relating the kth component to the qth response variable. Let η_{iq} and $g(\cdot)$ denote the ith linear predictors of the qth response and a link function, respectively. We assume that all the predictors and response variables are standardized with zero means and unit variances (Takane & Hwang, 2005). The GEE-ERA model is then expressed as:

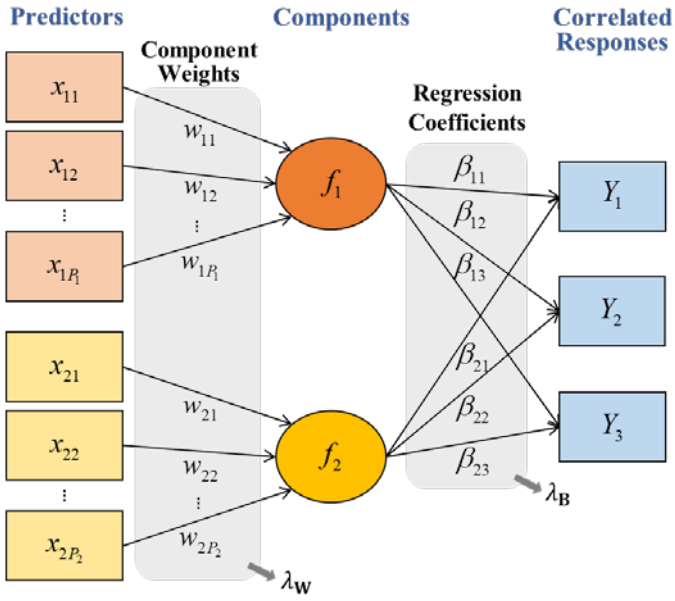


Figure 3.2.1. An example of GEE-ERA model. Square boxes indicate observed predictor and response variables. Circles represent predictor components. Two regularization parameters, $\lambda_{\mathbf{W}}$ and $\lambda_{\mathbf{B}}$, determine the strength of the regularization on component weights and regression coefficients, respectively

$$g(\mu_{iq}) = \eta_{iq} = \sum_{k=1}^{K}\left[\sum_{p=1}^{P_k} x_{ikp} w_{kp}\right]\beta_{kq} = \sum_{k=1}^{K} f_{ik}\beta_{kq} \qquad (3.2.1)$$

where the marginal expectation of the responses μ_{iq} is related to a linear predictor through a known link function. Figure 3.2.1 displays an example of the GEE-ERA model, where three response variables are assumed to be affected by each of the two components.

Let $\tilde{y}_i = [y_{i1}, \cdots, y_{iQ}]'$ be a Q by 1 vector of the responses of the ith respondent. Let Σ_i be the Q by Q within-respondent covariance matrix of $\tilde{y}_i$. When respondents are measured on multiple response variables simultaneously, the assumption of independence of response variables in ordinary ERA can be violated. Moreover, the true covariance structure is often unknown in practice. To resolve these issues in ERA, the method of GEE (Liang & Zeger, 1986) was applied to specify the unknown covariance structure using the so-called "working" correlation matrix. The working covariance matrix has the form

$$\text{cov}(\tilde{y}_i) = \Sigma_i = \phi A_i^{1/2} R_i(\boldsymbol{\alpha}) A_i^{1/2}, \quad (3.2.2)$$

where $R_i(\boldsymbol{\alpha})$ is a Q by Q working correlation matrix that is assumed to be fully specified by the vector of unknown nuisance parameters $\boldsymbol{\alpha}$, and $A_i^{1/2}$ is a Q by Q diagonal matrix of marginal variances with var(μ_{iq}) as the qth diagonal element (Liang & Zeger, 1986). Liang and Zeger (Liang & Zeger, 1986) suggested various choices for $R_i(\boldsymbol{\alpha})$ (see Section 3.3.2), which is constant across all respondents. In this way, we can treat the covariance structure as a nuisance instead of attempting to model it accurately when estimating ERA parameters. This method also can provide asymptotically unbiased parameter estimates and their robust standard errors regardless of the covariance structure specified (Lee et al., 2019).

3.2.2. ***Parameter Estimation and Significance Testing***

GEE-ERA aims to estimate both ERA parameters (i.e., w_{kp} and β_{kq}) and nuisance correlation parameters (i.e., $\boldsymbol{\alpha}$ and ϕ) in an iterative manner. Specifically, it seeks to minimize the following penalized least squares criterion for estimating parameters:

$$\varphi_{(\alpha,\mathbf{W},\mathbf{B})} = \sum_{i=1}^{N} [(\tilde{z}_i - \mathbf{B}'\mathbf{W}'\tilde{x}_i)'\Sigma_i^{-1}(\tilde{z}_i - \mathbf{B}'\mathbf{W}'\tilde{x}_i)] + \lambda_{\mathbf{W}}\text{trace}(\mathbf{W}'\mathbf{W}) + \lambda_{\mathbf{B}}\text{trace}(\mathbf{B}'\mathbf{B}) \quad (3.2.3)$$

where $\tilde{z}_i$ is a Q by 1 vector of the so-called adjusted response variable (McCullagh & Nelder, 1989, Chapter 2), $\mathbf{B}$ denotes a K by Q matrix of regression coefficients, $\mathbf{W}$ denotes a $P = \sum_{k=1}^{K} P_k$ by K matrix of component weights, $\tilde{x}_i$ denotes a vector of predictors for the ith respondent, and $\lambda_{\mathbf{W}}$ and $\lambda_{\mathbf{B}}$ denote tuning parameters for component weights and regression coefficients, respectively. The tuning parameters control the influence of the ridge penalty terms, trace($\mathbf{W}'\mathbf{W}$) and trace($\mathbf{B}'\mathbf{B}$). We apply G-fold CV to determine the values of $\lambda_{\mathbf{W}}$ and $\lambda_{\mathbf{B}}$ automatically. To minimize (3.2.3), GEE-ERA uses a regularized alternating least squares algorithm, in which each of $\mathbf{W}$, $\mathbf{B}$, and Σ_i is updated, with the other two parameter sets held

constant, until convergence. Refer to Appendix C. Parameter Estimation in GEE-ERA for a detailed description of the algorithm.

To test statistical significance of parameter estimates, GEE-ERA can use resampling methods, such as permutation tests for obtaining exact *p*-values (as described in Lee et al., 2019) and bootstrapping (Efron & Tibshirani, 1986) for constructing confidence intervals. In the present analysis, we used bootstrap percentile confidence intervals, i.e., the 2.5^{th} and 97.5^{th} percentiles of bootstrap distribution of parameter estimates based on 1,000 bootstrapped replications of the data.

3.3. An Empirical Application

3.3.1. ***Data and Model Specification***

The data used here is a subset of the 2012 National Survey on Drug Use and Health (NSDUH) dataset (United States Department of Health and Human Services, Substance Abuse and Mental Health Services Administration [SAMHSA], 2015). NSDUH has been conducted every year in all 50 states and the District of Columbia since 1971. The objective of this survey is to serve as a major source of information on tobacco, alcohol, and drug use, and on mental health and other health-related issues in the United States. The 2012 NSDUH was conducted from January through December 2012 and interviewed US residents aged 12 and older. Among 51 states, eight of them had a sample designed to yield 3,600 respondents per state, and the remaining 43 states had a sample designed to yield 900 respondents per state. The respondents were asked to answer various questions regarding their use of substances (e.g., tobacco, alcohol, illicit drugs, etc.), as well as mental and physical health issues. Each respondent's socio-demographic characteristics (e.g., age, race, marital status, education, financial circumstances, etc.) were also measured.

Table 3.3.1. A description of variables and summary statistics for the 2012 NSDUH data

Variable Names	Measures (Range or Categories)	Mean (Q1, Q3) [a]
Response Variables		
Y_1: Cigarettes	Number of cigarettes smoked per response in past month	200 (14, 315)
Y_2: Alcohol	Number of alcohol beverage drank in past month	55 (12, 64)
Y_3: Marijuana	On average, number of days used marijuana or hashish during the past 12 months	8.7 (2, 13)
Predictors		
$\mathbf{F}_1$: Age of first use		
Cigarette onset	Age of first use	15.81 (14, 18)
Alcohol onset	Age of first use	16.82 (15, 18)
Marijuana onset	Age of first use	16.94 (15, 18)
$\mathbf{F}_2$: Mental health		
Distress level	Nonspecific psychological distress scale (K6) score	2.01 (0, 2)
Impairment	Daily functional impairment due to problems with emotions, nerves, or mental health	1.09 (0, 3)
Suicidal thought	Serious thoughts of suicide in the past year (Yes=1/No=0)	%Yes: 9.58
Depression	Major depressive episode in the past year (Y=1/N=0)	%Yes: 12.5
$\mathbf{F}_3$: SES		
Education	5^{th} grade or less (=5), 6^{th} grade (=6), …, Sophomore/Junior (=14), Senior/Grad or more (=15)	12.41 (12, 14)
Insurance	Having any health insurance (Y/N)	%Yes: 71.75
Family income	Less than \$10,000 (=1), ~\$19,999 (=2), …, ~\$74,999 (=6), \$75,000 or more (=7)	4 (2, 6)
Employment Status	Employed (Y=1/N=0)	%Yes: 67.02

[a] For continuous variables, the first quartile (Q1), mean, and third quartile (Q3) are given.

In the present analysis, we examined the effects of predictors related to substance initiation age, mental health, and SES on cigarette, alcohol, and marijuana use. Table 3.3.1 presents summary statistics of all variables included in the analysis using data from $N = 881$ respondents. The three response variables, all referring to monthly use on average, are: the number of cigarettes smoked (Y_1), the number of alcoholic beverages consumed (Y_2), and the number of days of marijuana or hashish use (Y_3). We identified a total of 11 predictors that were available in the 2012 NSDUH data based on previous studies concerning the predictors of substance use on samples of US adults. Then, the predictors were grouped into the three

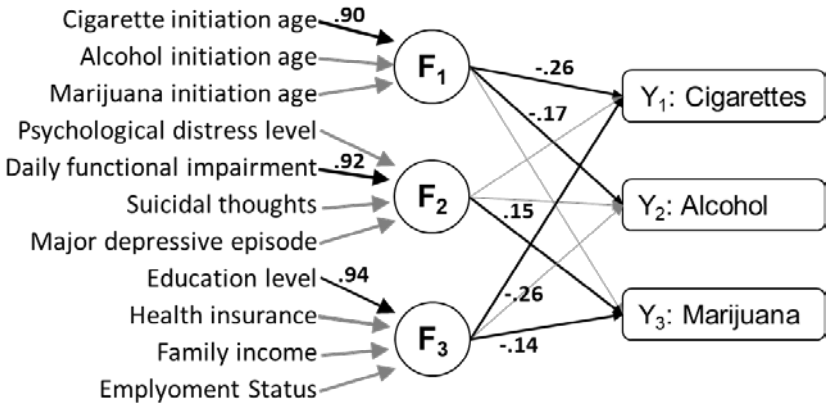


Figure 3.3.1. The specified ERA model for the 2012 NSDUH dataset. Black and bolded arrows represent statistically significant component weights and regression coefficients using bootstrap confidence intervals with $\lambda_{\mathbf{W}}$ =0.12 and $\lambda_{\mathbf{B}}$=0.

categories—substance initiation age (F_1), mental health (F_2), and SES (F_3)—which were represented as components in the ERA model. Table 3.3.1 also shows which component is associated with which predictors. Figure 3.3.1 displays the specified GEE-ERA model, where three sets of predictors related to F_1, F_2, and F_3 were to influence each of three response variables.

*3.3.2. **Working Correlation Structure of Substance Use Variables***

As noted above, previous studies suggested the co-occurrence of the three response variables. In the present data, there was a significant positive association between Y_1 and Y_2, r = .18, p < .01. Also, Y_1 and Y_3 were positively correlated, r = .16, p < .01, whereas Y_2 and Y_3 were not, r = -.02, p = .58.

The top row of Table 3.3.2 illustrates the four different working correlation structures considered in GEE-ERA to model the relationships in their co-occurrence: independent (all pairwise correlations fixed to zero), exchangeable (all correlations assumed to be equivalent), autoregressive or AR-1 (all first-order correlations assumed to be equivalent and higher-order correlations a function of the first-order correlation parameter), and unstructured (all

Table 3.3.2. The estimated working correlation and dispersion parameters across four different working correlation structures using the 2012 NSDUH data.

	Independent	Exchangeable	AR-1	Unstructured
Working correlation structures	$\begin{pmatrix} - & 0 & 0 \\ 0 & - & 0 \\ 0 & 0 & - \end{pmatrix}$	$\begin{pmatrix} - & \rho & \rho \\ \rho & - & \rho \\ \rho & \rho & - \end{pmatrix}$	$\begin{pmatrix} - & \rho & \rho^2 \\ \rho & - & \rho \\ \rho^2 & \rho & - \end{pmatrix}$	$\begin{pmatrix} - & \rho_1 & \rho_2 \\ \rho_1 & - & \rho_3 \\ \rho_2 & \rho_3 & - \end{pmatrix}$
Working correlation estimates	—	$\hat{\rho}$ =-.003	$\hat{\rho}$ =-.005	$\hat{\rho}_1$ =-.015, $\hat{\rho}_2$ =.001, $\hat{\rho}_3$ =.036
$\hat{\phi}$	.002	.002	.002	.002
QIC	2.853	2.853	2.860	2.869

Table 3.3.3. The estimated component weights for the GEE-ERA model in Figure 3.3.1 with different working correlation structures using the 2012 NSDUH data. Bolded numbers indicate statistically significant estimates using bootstrap confidence intervals.

		Working Correlation			
Components	Predictors	Independent	Exchangeable	AR-1	Unstructured
F$_1$: Age of first use	Cigarette onset	**.90**	**.90**	**.90**	**.90**
	Alcohol onset	.34	.34	.34	.33
	Marijuana onset	.02	.02	.02	-.01
F$_2$: Mental Health	Distress level	-.16	-.16	-.16	-.16
	Impairment	**.92**	**.92**	**.92**	**.92**
	Suicidal thought	.45	.45	.44	.44
	Depression	-.19	-.17	-.17	-.17
F$_3$: SES	Education	**.94**	**.94**	**.94**	**.94**
	Insurance	.28	.28	.27	.27
	Family income	-.09	-.09	-.08	-.08
	Employment Status	-.29	-.29	-.29	-.29

correlations assumed to be different and not systematically related). Table 3.3.2 also summarizes the working correlation and dispersion parameter estimates for each type of correlation structure from the present analysis, as well as the value of QIC, a modified Akaike information criterion for GEE models (Pan, 2001). All results in the table were obtained without any regularization, i.e., $\lambda_{\mathbf{W}} = \lambda_{\mathbf{B}} = 0$.

As shown in the table, the estimated correlation parameters changed in both sign and magnitude across the chosen correlation structures. However, the GEE-ERA parameter

Table 3.3.4. The estimated regression coefficients for the GEE-ERA model in Figure 3.3.1 with four different working correlation structures using the 2012 NSDUH data. Bolded numbers indicate statistically significant estimates using bootstrap confidence intervals.

			Working Correlation			
Components		Responses	Independent	Exchangeable	AR-1	Unstructured
$\mathbf{F}_1$: Age of first use	→	$\mathbf{Y}_1$: Cigarettes	**-.26**	**-.26**	**-.26**	**-.26**
		$\mathbf{Y}_2$: Alcohol	**-.17**	**-.17**	**-.17**	**-.17**
		$\mathbf{Y}_3$: Marijuana	-.09	-.09	-.08	-.08
$\mathbf{F}_2$: Mental Health	→	$\mathbf{Y}_1$: Cigarettes	.12	.12	.12	.12
		$\mathbf{Y}_2$: Alcohol	-.02	-.02	-.02	-.02
		$\mathbf{Y}_3$: Marijuana	**.15**	.15	.15	.15
$\mathbf{F}_3$: SES	→	$\mathbf{Y}_1$: Cigarettes	**-.26**	**-.26**	**-.26**	**-.26**
		$\mathbf{Y}_2$: Alcohol	-.05	-.05	-.05	-.05
		$\mathbf{Y}_3$: Marijuana	**-.14**	**-.14**	**-.14**	**-.14**

estimates in Table 3.3.3 and Table 3.3.4 were robust across different working correlation specifications. The final working correlation was chosen based on the value of QIC: Since independent and exchangeable structures resulted in equal QIC values, the more parsimonious of the two, i.e., independent, was chosen.

3.3.3. ***Regularization and Empirical Results***

After choosing the final correlation structure, we applied regularization on both component weights and regression coefficients. As the values of the regularization strengths, i.e., $\lambda_{\mathbf{W}}$ and $\lambda_{\mathbf{B}}$, are dependent on the data, they can be determined using data-driven methods, such as CV. We used 10-fold CV for different possible values of $\lambda_{\mathbf{W}}$ and $\lambda_{\mathbf{B}}$. The optimum values were chosen by comparing the average mean-squared errors, where the values ranged from 0 to 10 with a step size of .05. The lowest error was obtained with $\lambda_{\mathbf{W}} = .15$ and $\lambda_{\mathbf{B}} = 0$. The statistically significant estimates of component weights and regression coefficients with these final values are given in Figure 3.3.1.

As depicted in the figure and Table 3.3.3, the component weight estimate for cigarette initiation age was positive and statistically significant, indicating that cigarette

initiation age contributed to forming F_1, initiation of substance use, in explaining substance uses. Neither alcohol nor marijuana initiation age were statistically significant. For F_2, mental health status, only the level of daily functional impairment showed a statistically significant contribution. Finally, for F_3, socioeconomic status, only education level made a significant contribution to explaining the use of the three substances.

Figure 3.3.1 and Table 3.3.4 show the statistically significant regression coefficient estimates. First, the negative association between F_1 and both Y_1 and Y_2 indicated that a younger age of substance initiation was associated with an increased number of cigarettes smoked and alcoholic beverages consumed, with the effect appearing larger for cigarette use. Additionally, worse mental health status was associated with more days of marijuana use among American adults. There was no influence of mental health status either on cigarette or on alcohol use. Finally, American adults with lower levels of SES were found to report greater levels of both cigarettes smoked and days of marijuana use, where cigarette use was more strongly associated with SES level than marijuana use.

3.4. Conclusion

The present analysis applied GEE–ERA, a recently proposed extension to ERA, to data from the 2012 NSDUH survey on substance use. Substance use, including use of multiple substances, is prevalent in the American population and the source of numerous public health concerns. Additionally, substance use is known to involve multiple categories of predictors, including the predictor sets considered in the present analysis—initiation of substance use, mental health status, and socioeconomic status. We investigated the relationship of these predictors with cigarette, alcohol, and marijuana use. GEE-ERA permits the simultaneous analysis of the numerous predictors and multiple, correlated response variables by

simultaneously conducting data reduction and multivariate multiple regression while also modeling the correlation structure of the response variables. This method also employs ridge-type regularization to address potential overfitting, determining the strength of the regularization automatically through cross-validation, and conducts significance tests on ERA parameters (i.e., component weights and regression coefficients) using bootstrapping. The method thus protects against the common problems of multicollinearity among predictors, overfitting, and improper use of asymptotic statistical inference while producing easy-to-interpret parameter estimates.

The present analysis has demonstrated the utility of GEE–ERA while also providing insight on the phenomenon of substance use in the US. Nevertheless, there are several ways to expand upon the present analysis. First, given that the NSDUH is an annual survey, the present analysis should be replicated with data from subsequent years. Also, future studies should include additional predictors that have been found to significantly relate to substance use, such as personality characteristics (Hittner, Penmetsa, Bianculli, & Swickert, 2020) or sexual orientation discrimination (Evans-Polce, Veliz, Boyd, Hughes, & McCabe, 2019). Unfortunately, the NSDUH data did not include variables relevant to these factors. And finally, considering previous research that uncovered heterogeneous subgroups characterized by demographic covariates (e.g., gender or ethnicity), each of which yielded different effects of predictors on substance use, it will be worthwhile to further extend GEE-ERA to identify potentially heterogeneous subgroups of observations based on such covariates.

References

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

Efron, B., & Tibshirani, R. (1986). Bootstrap Methods for Standard Errors, Confidence Intervals, and Other Measures of Statistical Accuracy. *Statistical Science, 1*(1), 54–75. https://doi.org/10.1214/ss/1177013815

Evans-Polce, R. J., Veliz, P. T., Boyd, C. J., Hughes, T. L., & McCabe, S. E. (2019). Associations between sexual orientation discrimination and substance use disorders: differences by age in US adults. *Social Psychiatry and Psychiatric Epidemiology,* 1–10. https://doi.org/10.1007/s00127-019-01694-x

Hittner, J. B., Penmetsa, N., Bianculli, V., & Swickert, R. (2020). Personality and substance use correlates of e-cigarette use in college students. *Personality and Individual Differences, 152,* 109-605. https://doi.org/10.1016/j.paid.2019.109605

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

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

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

Liang, K. Y., & Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika, 73*(1), 13–22. https://doi.org/10.1093/biomet/73.1.13

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

National Institute on Drug Abuse. (2019). Drugged driving. *National Institutes of Health; U.S. Department of Health and Human Services.*, pp. 1–5. Retrieved from https://www.drugabuse.gov/publications/drugfacts/drugged-driving

Pan, W. (2001). Akaike's information criterion in generalized estimating equations. *Biometrics, 57*(1), 120–125. https://doi.org/10.1111/j.0006-341X.2001.00120.x

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

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

U.S. Department of Health and Human Services. (2004). The Health Consequences of Smoking: A Report of the Surgeon General. *National Library of Medicine, 2012*, 51576–51576. https://doi.org/10.1002/yd.20075

United States Department of Health and Human Services. Substance Abuse and Mental Health Services Administration. Center for Behavioral Health Statistics and Quality.

(2013). National Survey on Drug Use and Health Database. Inter-university Consortium for Political and Social Research (ICPSR) [distributor]. https://doi.org/10.3886/ICPSR35509.v1

Chapter 4. Prediction-Oriented Model Selection Metrics for Extended Redundancy Analysis

Abstract

In many areas of psychology, it is important to develop a model in such a way that it can predict health or behavioral outcomes (e.g., relapse of drug taking, symptoms of mental disorders, responses to treatment in new patients) well. An inherent challenge in building a prediction model is how to adequately assess the performance of the selected model on unseen data. A conventional way is to evaluate a model's performance in the sample used for model development, it is, however, well known that such *apparent* performance is an overly optimistic estimate of *true* prediction performance. As an alternative approach, in this chapter, I introduce several new metrics for evaluating the predictive ability of ERA models, focusing on their performance on so-called out-of-sample data that are not used for parameter estimation. Although considerable work has been done in statistics and machine learning in order to examine the utility of resampling methods (such as cross-validation and the bootstrap) for assessing such out-of-sample prediction, to date, no research has been carried out to apply these general tools to ERA. Thus, I conduct a simulation study to evaluate the relative performance of different out-of-sample prediction error estimators for ERA. This study may provide researchers with information on which error estimator is the best to find the true model when mis-specified (i.e., underfitted and overfitted) models are considered.

Keywords: Extended redundancy analysis, model selection, underfitted or overfitted models, out-of-sample prediction error, *k*-fold cross-validation, leave-one-out cross-validation, out-of-bag bootstrap, .632 bootstrap, .632+ bootstrap

4.1. Introduction

The present chapter concerns the assessment of the performance of ERA models, which is critical for model selection or development. As discussed in the previous chapters, one existing method of choice is to calculate FIT in (1.2.8), an overall goodness of fit measure for ERA (Takane & Hwang, 2005). Other measures of overall model fit for parametric ERA are based on penalized-likelihood criteria, such as AIC_{ERA} and BIC_{ERA}, which take model complexity into account (DeSarbo et al., 2015). All these existing metrics represent "in-sample" model evaluation metrics, in which the same dataset is used to develop the model and evaluate its performance. Naturally, this can lead to overly optimistic views of the model's performance: the more closely we fit the model to the training sample—a set of data used to estimate parameters, the better it will perform when being evaluated on the same sample. This is a well-known statistical phenomenon called "optimism" (Efron, 1983; Efron & Tibshirani, 1997; Hastie et al., 2009, Chapter 7).

When researchers are interested in predicting important health or behavioral outcomes to the benefit of the broader population, relying only on such "optimistic" in-sample model assessment metrics is not ideal because it provides little information about the model's performance on "out-of-sample". For example, in studies on cognitive impairment in older adults (Na, 2019; Choi & Jin, 2018), patient responses to treatments for depression (Cuijpers et al., 2013), and user response patterns in online advertising (Zhang, Du, & Wang, 2016), the goal of model development is to select a prediction model that can best assist practitioners with decision-making in unseen cases (e.g., treatment recommendation for new patients). To develop such models that can generalize beyond the current sample, researchers in the social and behavioral sciences would be better served by assessing the model based on out-of-sample performance metrics.

Thus, in this chapter, I introduce several new model evaluation metrics for ERA, each of which aims to quantify how well a model performs in out-of-sample data. But before discussing the out-of-sample metrics, the degree of optimism in existing in-sample model performance evaluation in ERA is briefly investigated. Hastie et al. (2009, Chapter 7) discussed, in general, the optimism of an in-sample model performance metric decreases linearly as the training sample size increases but increases with model complexity. Thus, a simulation study is carried out to examine the behavior of in-sample FIT and prediction error, focusing on how the degree of optimism is affected by varying training sample sizes and the number of predictors per component across different model specifications (e.g., over-specified models with additional parameters).

The next section will illustrate several strategies for assessing out-of-sample performance of ERA models to correct for the optimism of traditional in-sample metrics. One suggested (and commonly used) remedy for correcting the optimism is to approximate the model assessment step by sample-reuse or resampling, such as cross-validation (CV; (Geisser, 1975; Stone, 1974) and the bootstrap (Efron, 1979, 1983). The basic idea behind these methods is avoiding optimism by using non-overlapping data for the model development and evaluation. Although considerable work has been done in statistics and machine learning on the use of various CV and bootstrap methods for out-of-sample model assessment, to date, no research has applied these general tools to the ERA framework. Thus, I formulate multiple different out-of-sample prediction error estimators for ERA based on (1) *k*-fold CV (*k* = 3, 5, and 10), (2) leave-one-out CV (LOOCV), (3) out-of-bag (OOB) bootstrap, (4) .632 bootstrap, and (5) .632+ bootstrap, and carry out simulation studies to evaluate their relative behavior and predictive performance, and investigate which error

estimator is best for identifying the true model when mis-specified models are included as possible candidates under different simulation conditions.

The rest of this chapter is organized as follows. Section 4.2 provides a formal formulation for in-sample and out-of-sample prediction error estimators. The use of the abovementioned resampling methods in ERA is also illustrated. Section 4.3 shows the result of a series of simulation studies to examine how sample size and model complexity effect the performance of in-sample and out-of-sample prediction error estimators. In all simulations, underfitted and overfitted models are considered to illustrate the behavior of each error estimator in mis-specified settings. The final section summarizes the findings, provides a guideline for practitioners on which resampling approach may be favored under which condition, and discusses the limitation of the study.

4.2. Methods

4.2.1. ***Assessment of Predictive Performance***

Consider a continuous response variable Y that is related to a predictor matrix $\mathbf{X}$ by a statistical model f: $\mathbf{X} \rightarrow Y$. Then, $\hat{f}(\mathbf{X})$ denotes the predicted responses estimated from the observed training set $T = \{(\boldsymbol{x}_1', y_1), \cdots, (\boldsymbol{x}_i', y_i), \cdots, (\boldsymbol{x}_N', y_N)\}$, where y_i is the ith value of Y and $\boldsymbol{x}_i$ is a predictor vector for the ith observation. We assume that the observations in T are random samples from a distribution F. Let denote $L(Y, \hat{f}(\mathbf{X}))$ the loss function[5] for measuring errors between Y and $\hat{f}(\mathbf{X})$. For example, in many regression-based models, a common choice for a continuous Y is the squared error loss, i.e., $L(Y, \hat{f}(\mathbf{X})) = (Y - \hat{f}(\mathbf{X}))^2$.

[5] The term *loss* in mathematical optimization is used to describe how much a model is losing compared to having made perfect predictions

As shown in (1.2.3) of Chapter 1, ERA also defines its objective function in terms of the mean squared error loss and seeks to minimize it over all possible parameter values.

Conventional goodness of fit measures for ERA inform how well a model fits to the training set, thus being of little utility when assessing a model's prediction capability. There are various loss functions to quantify the overall predictive performance, but root mean square error (RMSE),

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{f}(\boldsymbol{x}_i))^2}\,, \tag{4.2.1}$$

is a reasonable choice for ERA, considering that all ERA parameters are estimated to minimize the mean squared error loss, i.e., the sum of squared prediction errors, as discussed above.

In predictive modeling, we wish to obtain a model that not only performs well on the training data, but also on independent unseen data. Thus, understanding how to estimate the error rate of a model when it is used to predict the future responses is important as it guides the choice of final model. Let $(\boldsymbol{x}_0', y_0)$ is a new independent test sample randomly drawn from *F*. The *true test error* or *generalization error* (Efron, 1983; Efron & Tibshirani, 1997; Hastie et al., 2009, Chapter 7), is the prediction error for $(\boldsymbol{x}_0', y_0)$,

$$\text{Err}_T = E_{(\boldsymbol{x}_0', y_0)}[L(y_0, \hat{f}(\boldsymbol{x}_0)) \mid T]. \tag{4.2.2}$$

Note that, in (4.2.2), only $(\boldsymbol{x}_0', y_0)$ is random with *T* being fixed, meaning that the true test error refers to the conditional error for the particular training set *T*. In practice, it is more amenable to estimate a model's prediction error as the expectation of Err_T (Efron, 1983; Efron & Tibshirani, 1997; Hastie et al., 2009, Chapter 7),

$$\text{Err} = E[\text{Err}_T] = E[L(Y, \hat{f}(\mathbf{X}))], \tag{4.2.3}$$

where everything random is averaged over. In many machine learning applications, where a large independent test set is available, the goal of model selection is to find a model that gives minimum expected test error in (4.2.3).

In the absence of a large independent test set, the simplest way to estimate Err_T is to use the *training error* or *apparent error*, defined by the average loss over the training data,

$$\mathrm{Err}_{\mathrm{Train}} = \frac{1}{N}\sum_{i=1}^{N} L(y_i, \hat{f}(\boldsymbol{x}_i)). \qquad (4.2.4)$$

As discussed previously, $\mathrm{Err}_{\mathrm{Train}}$ is typically smaller than Err_T, i.e., $\mathrm{Err}_{\mathrm{Train}}$ tends to be biased downward as an estimate of Err_T because the same observations are used twice, both for fitting $\hat{f}(\mathbf{X})$ and for evaluating the prediction error of $\hat{f}(\mathbf{X})$ (Efron & Tibshirani, 1997). To alleviate this inherent optimism in $\mathrm{Err}_{\mathrm{Train}}$, various resampling methods can be employed.

4.2.2. ***Resampling Methods for Out-of-Sample Model Assessment***

All prediction error estimators introduced in this section aim to estimate Err_T more accurately than the apparent error by adopting different resampling methods. Key references on the use of different variants of CV and the bootstrap for prediction error estimation are Breiman and Spector (1992), Efron (1983), and Efron and Tibshirani (1997).

A straightforward approach for correcting the optimism in $\mathrm{Err}_{\mathrm{Train}}$ is to randomly split the observed data *T* in two parts: one for developing the model (*training* or *learning set*) and the other for measuring its predictive performance (*validation set*). With this split-sample approach, model performance is determined on independent data not used for model development. However, there are two criticisms of this procedure. First, it is inefficient, especially when the size of *T* is small, owing to its reduction of the size of both learning and validation sets. Second, high variability in the estimated predictive performance can be introduced because of its reliance on a single split of *T*. Thus, *k*-fold CV, which can be

considered an extension of the split-sample approach, is preferred. This method randomly assigns *N* observations to one of *k* partitions such that the partitions are of nearly equal size. Subsequently, the learning set contains all but one of the partitions which is labeled the validation set. We fit the model to the learning set, and calculate the prediction error (e.g., RMSE) of the fitted model to the validation set. After repeating this for all *k* folds, the *k* prediction error estimates are averaged, resulting in the *k*-fold CV estimate of prediction error, $\overline{\text{Err}}_{(\text{cv},k)}$. LOOCV is the most extreme case of *k*-fold CV, where the number of folds equals the number of observations (i.e., $k = N$) and each observation is individually assigned to the validation set.

Efron (1983) proposed and compared a number of bootstrap resampling variants for the assessment of a model's predictive performance, which are generally referred to as *out-of-bag* (OOB) estimators in the statistics and machine learning literature. Calculating the OOB prediction error begins with bootstrap sampling. Let *B* be the number of bootstrap replications. For each draw, a bootstrap sample contains only 63.2% of the original data on average (referred to as *in-bag sample*) due to the sampling with replacement. The prediction error is assessed on the remaining 37% of the data (out-of-bag data) for each bootstrap draw and subsequently averaged over the *B* iterations, resulting in the OOB estimate of prediction error, $\overline{\text{Err}}_{(\text{OOB})}$. There are two more variations of the OOB estimator: the .632 estimator, $\overline{\text{Err}}_{(.632)}$, and the .632+ estimators, $\overline{\text{Err}}_{(.632+)}$. Both aim to correct the underestimated $\text{Err}_{\text{Train}}$ as a weighted combination of $\text{Err}_{\text{Train}}$ and $\overline{\text{Err}}_{(\text{OOB})}$: $\omega \cdot \text{Err}_{\text{Train}} + (1-\omega) \cdot \overline{\text{Err}}_{(\text{OOB})}$. The value of ω

is fixed as .632 for the .632 estimator, whereas ω is determined based on the so-called "no-information error rate" for the .632+ estimator (Efron and Tibshirani, 1997)[6].

4.3. Simulation Study

In this study, the following experimental factors were considered: (1) sample size for training data (T), N = 50, 100, 200, 500, and 1,000, and (2) the number of predictors per component, N_p = 2, 4, 6, and 8. Using the data generation procedure described in Section 2.3.1, simulation data were generated for 20 different scenarios (5 different sample sizes x 4 different numbers of indicators). For each scenario, different mis-specified ERA models were considered to see whether the true model was chosen based on different model assessment metrics. More specifically, four model specifications were considered: (1) under-specified model, f0: $y_i = b_1 f_{i1} + e_i$, (2) correctly-specified model (i.e., data generating model), f1: $y_i = b_1 f_{i1} + b_2 f_{i2} + e_i$, (3) over-specified model with a component interaction term, f2: $y_i = b_1 f_{i1} + b_2 f_{i2} + b_3 (f_{i1} \cdot f_{i2}) + e_i$, and (4) over-specified model with interaction and quadratic terms, f3: $y_i = b_1 f_{i1} + b_2 f_{i2} + b_3 (f_{i1} \cdot f_{i2}) + b_4 (f_{i1}^{\,2}) + b_5 (f_{i2}^{\,2}) + e_i$. In all conditions, the total number of repetitions was 1,000.

4.3.1. ***Optimism in In-Sample Measures***

Figure 4.3.1 shows the apparent performance of FIT (i.e., the in-sample performance of FIT) for four different ERA models (f0, …, f3) in relation to sample size (N) and the number of

[6] In brief, the weight is dependent on the relative amount of overfitting coefficient R: $\omega = .632/(1-.368 \cdot R)$. The relative overfitting R is large when the difference between $\text{Err}_{\text{Train}}$ and $\overline{\text{Err}}_{(\text{OOB})}$ is relatively large. In this case, R and ω approach 1, indicating that the estimated prediction error is largely based on $\overline{\text{Err}}_{(\text{OOB})}$. When the overfitting is small, R approaches 0 and ω .632, resulting in similarity between the .632 and .632+ estimators.

indicators per component (N_p). Boxplots were constructed to show the distribution of estimated FIT values over 1,000 repetitions. Each dotted line indicates the Err estimate in (4.2.3) which was estimated over 1,000 simulated independent test sets of size N each. As the boxplots show, with large sample size and smaller number of predictors, the median apparent performance of FIT (boxplot centers) approached the test performance (dotted lines). We also note a reduction in the variability of the model performance estimates (the length of boxplots) in such conditions. For all simulation conditions, however, the true model (f1) was never selected. In addition, the median apparent performance was always above the dotted line, indicating the optimism of conventional FIT as an estimate of true predictive performance. For the overfitted models, f2 and f3, the apparent performance of FIT tended to reach its maximum value (i.e., 1) rapidly in small samples and many predictors, showing that the optimism in the apparent FIT measure can be problematic when overfitting may be an issue in model selection.

In Figure 4.3.2, boxplots show the apparent error measured based on RMSE. The differences between Figure 4.3.1 and Figure 4.3.2 are minimal in terms of the optimism of in-sample model assessment. But, by looking at the variability of the estimates across different model specifications (f0, …, f3), we can see that RMSE is less affected by model mis-specification. For example, in Figure 4.3.2, the difference in variabilities of apparent RMSE for f0, f1, f2, and f3 became very minimal when the sample size increased (e.g., $N \geq 100$), while the variability of apparent FIT for the true model (f1) in Figure 4.3.1 was always the largest for all simulation condition compared to that of each mis-specified model.

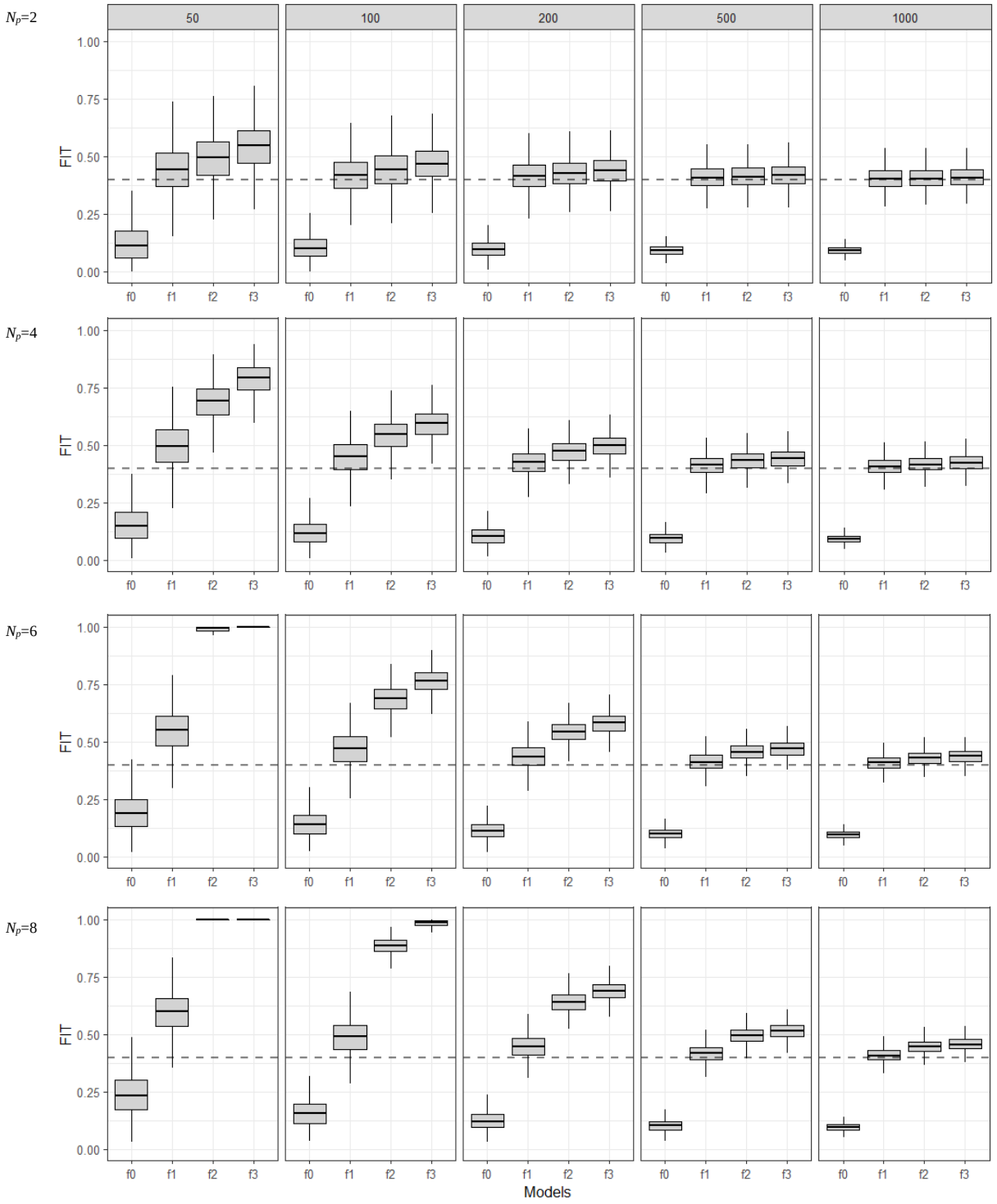


Figure 4.3.1. Behavior of in-sample FIT values as the training set sample size (N=50,…,1000) and model complexity (Np=2,…,8) are varied for correctly- and incorrectly-specified ERA models (f0, f1, f2, and f3).

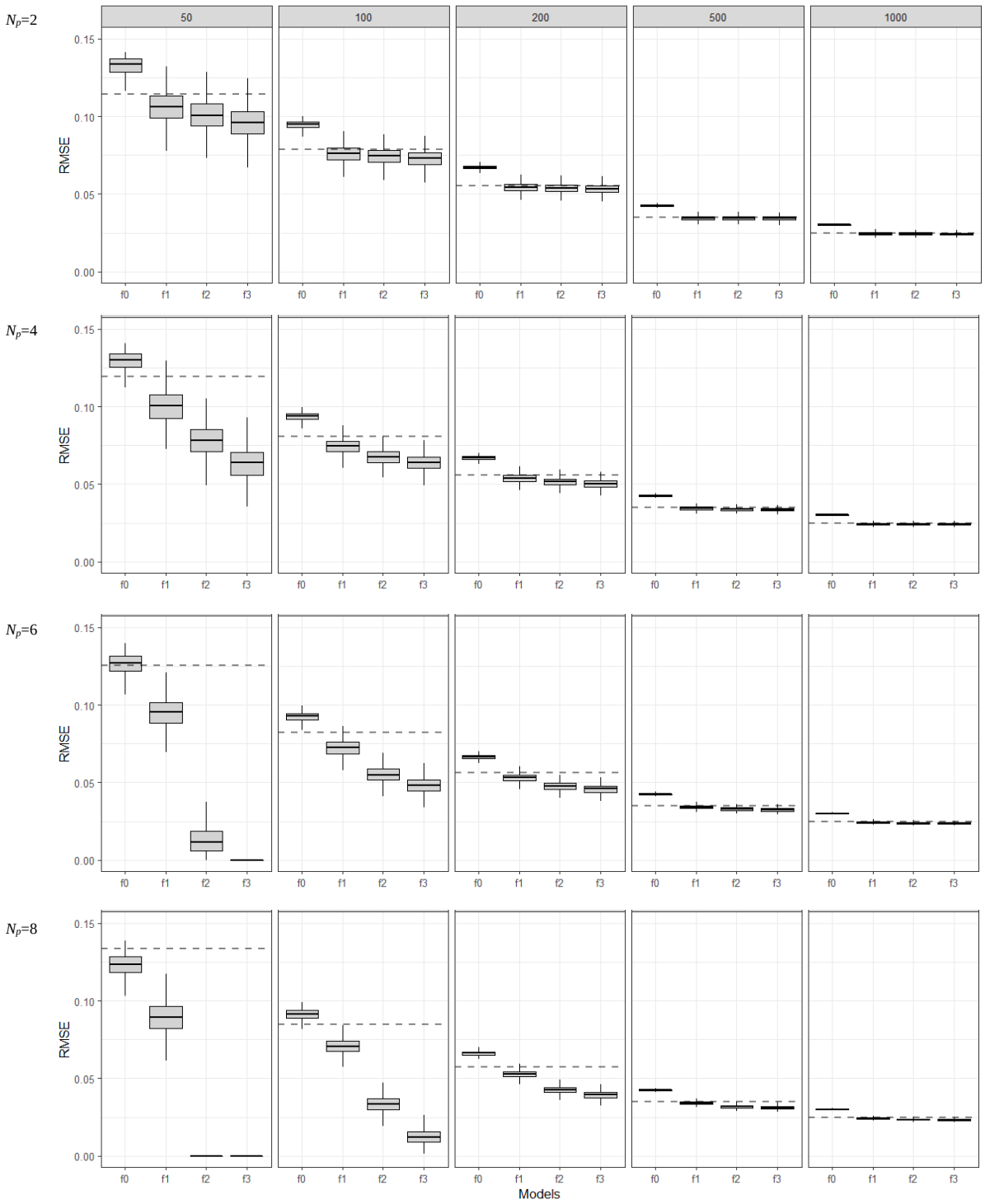


Figure 4.3.2. Behavior of apparent RMSE as the training set sample size (*N*) and model complexity (*Np*) are varied for correctly- and incorrectly-specified ERA models (f0, f1, f2, and f3).

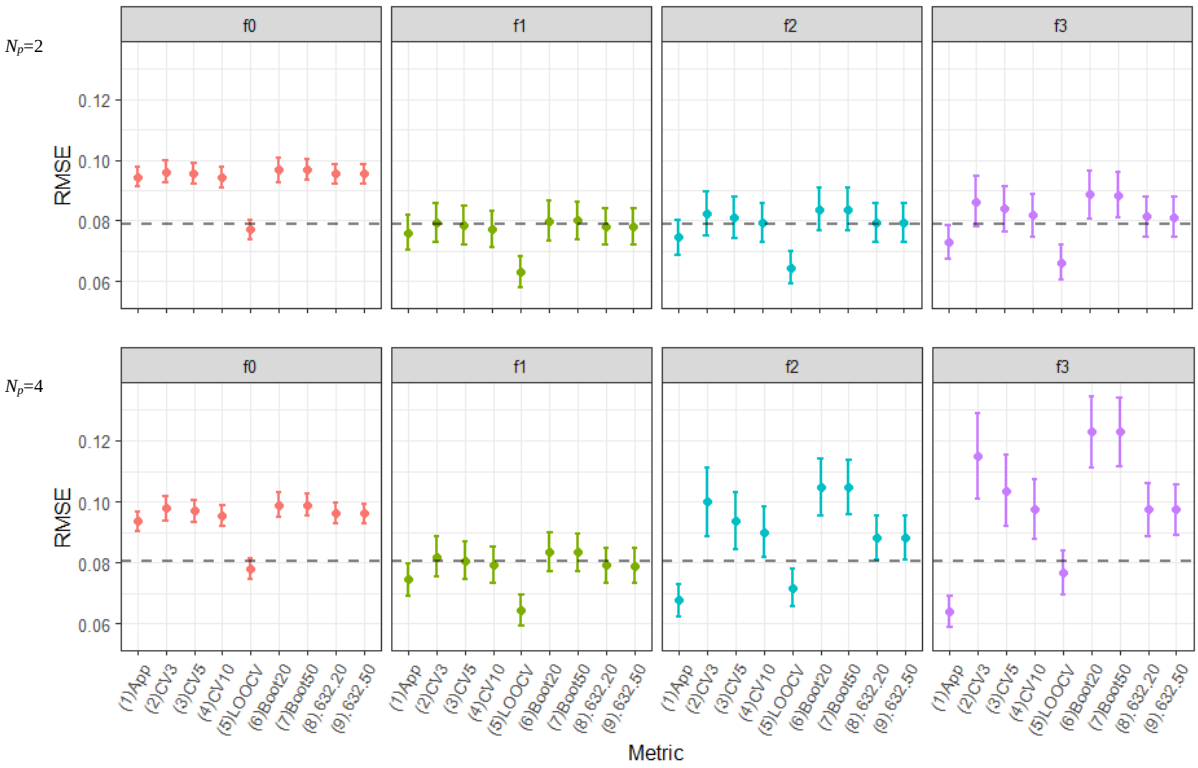


Figure 4.3.3. Behavior of different out-of-sample prediction error estimators based on various resampling methods, as the model complexity (*Np* = 2 and 4) are varied for correctly- and incorrectly-specified ERA models (f0, f1, f2, and f3). The training set sample size is *N* = 100.

4.3.2. ***Behavior of Out-of-Sample Estimators***

In this section, all results are discussed but only a limited number of figures are displayed because the differences in the error estimators across the resampling methods were minimal as the sample size increases, *N* > 100. The full compilation of simulation results archived on the author's GitHub at https://github.com/QuantMM.

Figure 4.3.3 displays the behavior of different prediction error estimators based on various resampling methods for correctly- and incorrectly-specified ERA models (f0, f1, f2, and f3), when *N* = 100 and *Np* = 2 and 4. The error is calculated based on RMSE for the apparent error (denoted by (1)App in the figure), *k*-fold CV estimators (for *k* = 3, 5, and 10; (2)CV3, (3)CV5, and (4)CV10, respectively), LOOCV estimator ((5)LOOCV), the regular

OOB bootstrap estimators $\overline{\text{Err}}_{(\text{OOB})}$ with B = 20 and 50 ((6)Boot20 and (7)Boot50), the .632 estimator with B = 20 and 50 ((8).632.20 and (9).632.50). Due to space limitations, the results of .632+ estimator are not included in the figure as there was no noticeable difference between the .632 and .632+ estimators in all simulation conditions. Also, the results obtained from $\overline{\text{Err}}_{(\text{OOB})}$ and the .632 estimators with B = 100 are not displayed because there was minimal improvement over those with B = 50. In the figure, each error bar represents one standard deviation of the expected value of the estimated errors over 1,000 repetitions. Each dotted line indicates the Err estimate in (4.2.3) which is estimated over 1,000 simulated independent test sets of size N each.

Most noticeably, all resampling estimators resulted in the smallest error for f1, thus the true ERA model was always selected across all simulation conditions. In this true model condition, all estimators (except the LOOCV estimator) successfully corrected the underestimated prediction error in the apparent error estimator and exhibited similar variabilities (the length of error bars). The LOOCV estimator had the lowest variability but showed noticeable downward bias. Additionally, the 5-fold CV estimator had the smallest bias, followed by the .632 and .632+ bootstrap methods with B = 20.

The behavior of error bars in the mis-specified conditions, i.e., f0, f2 and f3, clearly shows that the bias and variance of each error estimator depends on the apparent performance of a model. When a model was underspecified, thus showing poor apparent performance, all error estimators overly overestimated the true prediction error but always had low variability. When overfitting occurred (f2 and f3), the k-fold CV estimators highly overestimated prediction error with high variability, which tended to be worse for a smaller value of k. The LOOCV estimator corrected such overly upward biased prediction error and had the smallest variability. This indicates that, when there is little bias in the k-fold CV estimators (as in the

true model condition, f1), downward bias can occur with the LOOCV estimator. This is possible because each learning set in the LOOCV procedure is very similar to the full observed data *T*. Similarly, the OOB bootstrap estimators showed large biases for overfitted models, where the upward bias was substantially reduced by the .632 and .632+ estimators. However, the advantage of increasing *B* from 20 to 50 was minimal.

4.4. Discussion and Recommendations

To build a model that generalizes the result beyond the current sample, especially when overfitting may be an issue, the use of out-of-sample model assessment metrics is crucial in model selection. There has been no discussion in the ERA literature as to out-of-sample error estimation for model selection, and no comparison of widely-used resampling methods has been performed to date. Thus, this chapter discussed several resampling strategies to estimate prediction error in the absence of independent future data, as alternatives to the conventional in-sample goodness of fit measures.

Simulation results demonstrated that the optimism of conventional in-sample model evaluation metrics was negligible in large samples (e.g., $N \geq 500$), but never disappeared. This implies that comparing two or more candidate models relying only on conventional FIT or in-sample RMSE is not recommended because these model evaluation metrics always favor more complex models (with a larger number of predictors per component and/or overfitted model), thereby being unable to select model resulting in a reproducible conclusion for future data.

The simulation study also highlighted the advantage of adopting CV and bootstrap methods to avoid overly optimistic assessment of a model's predictive performance. Some general conclusions may be summarized as follows. Firstly, the differences among the

resampling methods (in terms of both bias and variance of estimators) decrease as the sample size increases, e.g., $N \geq 200$, even for mis-specified models. Secondly, when highly complex models are considered, the *k*-fold CV with smaller number of folds and the regular OOB bootstrap methods may perform poorly compared to other resampling methods. Thirdly, for largely over-specified models, the LOOCV estimator was a reasonable choice as it resulted in the lowest bias and variability. Fourthly, $B = 20$ bootstrap replications would be sufficient for the regular OOB bootstrap estimator and its variants. The advantage of increasing *B* from 20 to 100 was minimal in terms of the variability of estimators. Lastly, overall, the .632 and .632+ estimators outperformed other estimators, but the 10-fold CV prediction error estimate approximated those of .632 and .632+ in almost all settings. Thus, for computationally burdensome analyses, 10-fold CV may be preferable over the OOB bootstrap estimators.

This chapter discussed the assessment of prediction performance in terms of RMSE error. Thus, the simulation results can be widely applicable for other ERA models for continuous responses fit by expected squared error loss. However, when response variables are discrete—which is often termed *classification problems* in statistics and machine learning, the best choices of resampling methods may differ substantially. Simulation studies on prediction error estimation in classification problems (e.g., Efron, 1997; Hastie et al., 2009, Chapter 7.3) demonstrate that the bias and variance of expected test error in (4.2.3) behave considerably differently for classification loss functions (e.g., 0-1 loss, the negative binomial log-likelihood known as *deviance* or *cross-entropy*) than they do for squared-error loss. Thus, future work is needed to explore the behavior of the resampling methods for estimating the expected test error in classification problems, especially focusing on the

effects of different loss functions and number of classification classes in each response variable.

In addition, further studies are necessary to examine the utility of prediction error assessment based on resampling methods in a wide range of real data applications. For example, the empirical application examples discussed in the previous chapters used public data collected from the 2012 National Survey on Drug Use and Health (NSDUH). The empirical applications were conducted somewhat in an exploratory fashion, i.e., there were many "researcher degrees of freedom" (Simmons, Nelson, & Simonsohn, 2011) to decide what predictors to include or exclude and which component sets to use. For investigation of generalizability of the results, possible candidate models with different levels of model complexity can be constructed and compared based on the model selection metrics discussed in this chapter. Further replication analyses using a few more NSDUH surveys from 2013 to 2018 to validate the findings from 2012 may help to reveal the utility of out-of-sample prediction error estimators, and at the same time, offer the promise of reducing overly optimistic assessment of model performance.

References

Breiman, L., & Spector, P. (1992). Submodel Selection and Evaluation in Regression. The X-Random Case. *International Statistical Review / Revue Internationale de Statistique, 60*(3), 319. https://doi.org/10.2307/1403680

Choi, H., & Jin, K. H. (2018). Predicting cognitive decline with deep learning of brain metabolism and amyloid imaging. *Behavioural Brain Research, 344*, 103–109. https://doi.org/10.1016/j.bbr.2018.02.017

Cuijpers, P., Berking, M., Andersson, G., Quigley, L., Kleiboer, A., & Dobson, K. S. (2013). A meta-analysis of cognitive-behavioural therapy for adult depression, alone and in comparison with other treatments. *Canadian Journal of Psychiatry, 58*(7), 376–385. https://doi.org/10.1177/070674371305800702

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

Efron, B. (1979). Bootstrap Methods: Another Look at the Jackknife. *The Annals of Statistics, 7*(1), 1–26. https://doi.org/10.1214/AOS/1176344552

Efron, B. (1983). Estimating the Error Rate of a Prediction Rule: Improvement on Cross-Validation. *Journal of the American Statistical Association, 78*(382), 316. https://doi.org/10.2307/2288636

Efron, & Tibshirani. (1997). Improvements on Cross-Validation: The 632+ Bootstrap Method. *Journal of the American Statistical Association, 92*(438), 548–560. https://doi.org/10.1080/01621459.1997.10474007

Geisser, S. (1975). The Predictive Sample Reuse Method with Applications. *Journal of the American Statistical Association, 70*(350), 328. https://doi.org/10.2307/2285815

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

Na, K. S. (2019). Prediction of future cognitive impairment among the community elderly: A machine-learning based approach. *Scientific Reports, 9*(1), 1–9. https://doi.org/10.1038/s41598-019-39478-7

Stone, M. (1974). Cross-Validatory Choice and Assessment of Statistical Predictions. *Journal of the Royal Statistical Society, 36*(2), 111–147. https://doi.org/10.2307/2984809

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

Zhang, W., Du, T., & Wang, J. (2016). Deep Learning over Multi-field Categorical Data. In *Advances in Information Retrieval. ECIR 2016. Lecture Notes in Computer Science* (Ferro N. et al., Vol. 9626, pp. 45–57). Springer, Cham. https://doi.org/10.1007/978-3-319-30671-1_4

Chapter 5. Conclusion

5.1. Summary of Results and Contributions

Psychologists are making serious effort to capture the complexities of human behavior and health issues, which naturally encourages greater use of a large number of predictors. For example, substance use is associated with a number of different categories of predictors, including an individual's mental health, mental disorders, physical health, quality of life, social conditions, and SES, to name a few. This easily gives rise to several tens of variables. ERA is especially efficient in such settings as it provides a simpler interpretation of predictor-response relationships by summarizing multiple sets of predictors into a new set of lower-dimensional components. Using domain-specific knowledge concerning which predictors are to be put together within a researcher-defined component facilitates the interpretability of components. The final model ensures predictability as well because ERA searches for components that maximize predictive accuracy, without having to eliminate any predictors of interest to avoid multicollinearity. On the top of that, the potential and practical usefulness of ERA lies in its predictive nature; as discussed in previous chapters, ERA has been well blended with many statistical techniques for regression problems, including generalized linear model, generalized estimating equations, regularization techniques, and various supervised machine learning algorithms.

The present research makes methodological contributions to expanding conventional ERA for the analysis of: (1) potential heterogenous subgroups of observations characterized by combinations of auxiliary covariates and (2) multiple correlated response variables when the assumption of independent observations is violated. A large public dataset concerning drug use among US residents was used to illustrate the empirical usefulness of these novel

extensions. In addition, the present research proposes new prediction-oriented model selection strategies for ERA based on out-of-sample model evaluation metrics. Each of the novel approaches to ERA presented in this dissertation adopts machine learning algorithms (such as recursive partitioning and various regularization techniques), commonly used statistical modeling frameworks (such as generalized estimating equations and resampling methods), or a combination of both. By doing so, all the proposed ERA extensions attempt to provide a *better* ERA solution—one which allows the amount of model flexibility necessary for adequate data fit while providing interpretable results that can be easily understood by domain experts and not only by quantitative researchers.

As noted earlier, numerous health studies based on US nationwide survey datasets suggest that certain groups of US residents are dissimilar to others with respect to socio-demographic and health characteristics. A standard way to explore such patterns of heterogeneity is to compare a group of observations with a particular covariate characteristic to another group with a different characteristic (e.g., a covariate, gender; females vs. males). However, it is difficult to know which covariates should be used and how those covariates interact with each other. There will also be increased complexity in both analysis and interpretation, particularly when a number of continuous and/or categorical covariates are of interest. Unlike this standard way of group comparisons, the adaptation of a recursive partitioning method in MOB-ERA allows an automatic detection of a meaningful combination of covariates for a specified ERA model, thereby being able to capture unknown but important covariate-dependent heterogeneity in a data-driven manner. The proposed method is not overly computationally burdensome (compared to conventional ERA) because a series of parameter instability tests for each partitioning covariate is performed based on the empirical score contributions that are already obtained when estimating parameters. In

addition, the resulting flowchart-like tree diagram facilitates an understanding of hierarchically nested covariate structures selected during data partitioning, displaying how the whole dataset is split into heterogeneous subgroups

The second extension of ERA discussed in Chapter 3 concerns data analytic issues stemming from correlated responses, such as clustered data or time-dependent repeated measures. The proposed method, ERA-GEE, combines ERA with penalized GEE to simultaneously analyze multiple correlated response variables that are affected by a common set of predictors, relaxing the assumptions of correct specification of the covariance structure of the responses. ERA-GEE offers two additional practical advantages of employing GEE: (1) the model can handle various types of responses, such as scale, binary, counts, events-in-trials, or any combinations of these and (2) does not require balanced design or equally spaced measurements for responses. The proposed method also entails the advantages of multivariate analysis, including the ability to provide more statistical power compared to conducting a series of univariate analyses, as well as the capability to glean a more holistic picture than looking at a single response at a time. The proposed method was successfully applied to the analysis of rare genetic variants that are associated with multiple metabolic syndrome measures (Lee et al., 2019), as well as to the study of co-occurring recreational substance use among US adults (Kim et al., in press).

In many psychology studies, it is often assumed that a sample at hand (i.e., a training set of data) is a good reflection of what will be encountered in future data; thus, the final model is selected as the one optimized in the training data. Comparing two or more candidate models based on such goodness-of-fit (GOF) assessment is not ideal because GOF model evaluation metrics always favor more complex models (which fit the training data too tightly), thereby limiting the generalizability of the selected model. But estimating model

performance in future samples relying only on information contained in the current sample is also a hard problem. To respond to this, the last section of this thesis discussed the use of computer-intensive resampling methods, including variants of cross-validation (CV) and the bootstrap, in order to provide new ways of assessing generalizability of ERA models. In fact, owing to improvements in statistical computing over the past years, it has become substantially easier to execute various resampling methods on modern laptops without much computational burden. A series of simulation studies illustrated that, over a wide range of different model complexities and sample sizes in correctly- and incorrectly-specified model conditions, all of the out-of-sample prediction error estimators favored the true model with the highest frequency. The estimators based on 10-fold CV, .632, and .632+ methods outperformed other resampling strategies, but the difference between error estimators became unnoticeable in large samples. As the first initiative in investigating out-of-sample model performance of ERA, the broader goal of this study is to bring this discussion into the field of psychology so that such predictive model assessment metrics can be effectively utilized for investigation of reproducibility and generalizability of psychological and behavioral data science.

5.2. Future Research Directions

This thesis presents three important contributions to the advancement of ERA. The limitations and future research directions are discussed in detail in each chapter, but they can be summarized as follows:

- (Chapter 2. MOB-ERA) It is well known that the obtained results from any types of recursive partitioning methods are likely to be inflexible when it comes to new samples even after applying various regularization techniques. In order to deal with

such high variability problem, I plan on advancing MOB-ERA from a "single tree" to a "forest" by applying bootstrap aggregating or bootstrap smoothing (also known as *bagging*; Breiman, 1996). A MOB-ERA-forest will provide a more stable and less overfitted result, where each MOB-ERA-tree is built upon a random bootstrap sample of the original data.

- (Chapter 3. GEE-ERA) Missing data are a common problem in many studies dealing with multiple grouped responses, especially in longitudinal studies. To obtain unbiased GEE estimates, the assumption on the pattern of missingness is missing completely at random (MCAR). A more refined GEE-ERA should be further studied to handle a specific pattern of missingness such as monotone dropout (i.e., when study subjects are fully observed up to a certain point but have no measurements at subsequent points) or non-monotone (intermittent) dropout. For this, various multiple imputation (MI) methods, e.g., an MI approach assuming a multivariate normal distribution or an adaptation of the fully conditional specification (FCS) with the use of Gibbs sampling, can be considered.
- (Chapter 4. Predictive Performance Assessment in ERA) Optimistic estimation of model performance in classification problems should be further examined. A variety of loss functions for prediction error can be considered, such as 0-1 loss and deviance. Especially, understanding the effect of the number of events per variable (EPV), instead of simple sample sizes, is important when binary or multinomial response variables are considered.

Moving forward, future work should consider how to capture important between-person differences in ERA or individual-specific effects, such as random regression coefficients (e.g., pattern of changes in health or behavioral outcomes varying across individuals) and

random intercepts (e.g., baselines of such outcomes across individuals). Within the GEE framework discussed in GEE-ERA, the interest was in controlling for some degree of dependence from correlated responses when estimating the marginal effects of multiple predictors over all observations. In this circumstance, the dependence among the responses is treated as a nuisance parameter that is not of direct interest. Moreover, the dependency structure is assumed to be the same for all observations. Thus, I plan to extend ERA into the multilevel modeling framework to investigate individual-specific effects while controlling for potential predictor effects. The motivation of this proposal stems from the analysis of the University of Michigan Health and Retirement Study (HRS) data, which is a national longitudinal study for investigating the socio-psychological characteristics and physiological states of older Americans in relation to their cognitive decline over time. The respondents who participated in the HRS survey were not all measured at the same initial ages, meaning that there is not a one-to-one correspondence between time points and age at time of measurement. Moreover, the age-related declines in cognitive functioning randomly vary across respondents. This in turn indicates that ERA needs to adopt a more flexible framework to properly capture the differences in baselines of age-related changes, and at the same time, to investigate whether age-related declines in cognitive functioning randomly vary across respondents.

References

Alin, A. (2009). Comparison of PLS algorithms when number of objects is much larger than number of variables. *Statistical Papers*, *50*(4), 711–720. https://doi.org/10.1007/s00362-009-0251-7

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

Bentler, P. M., & Chou, C.-P. (1987). Practical Issues in Structural Modeling. *Sociological Methods & Research*, *16*(1), 78–117. https://doi.org/10.1177/0049124187016001004

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. (1989). *Structural Equations with Latent Variables*. Hoboken, NJ, USA: John Wiley & Sons, Inc. https://doi.org/10.1002/9781118619179

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

Bollen, K. A., & Davis, W. R. (2009). Causal Indicator Models: Identification, Estimation, and Testing. *Structural Equation Modeling: A Multidisciplinary Journal*, *16*(3), 498–522. https://doi.org/10.1080/10705510903008253

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.

Breiman, L., & Spector, P. (1992). Submodel Selection and Evaluation in Regression. The X-Random Case. *International Statistical Review / Revue Internationale de Statistique*, *60*(3), 319. https://doi.org/10.2307/1403680

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

Cheung, M. W. L., & Jak, S. (2016). Analyzing big data in psychology: A split/analyze/meta-analyze approach. *Frontiers in Psychology*, *7*, 1–13. https://doi.org/10.3389/fpsyg.2016.00738

Choi, H., & Jin, K. H. (2018). Predicting cognitive decline with deep learning of brain metabolism and amyloid imaging. *Behavioural Brain Research*, *344*, 103–109. https://doi.org/10.1016/j.bbr.2018.02.017

Choi, Kyung, M., Hwang, H., & Park, J. H. (2020). Bayesian Extended Redundancy Analysis: A Bayesian Approach to Component-based Regression with Dimension Reduction. *Multivariate Behavioral Research*, *55*(1), 30–48. https://doi.org/10.1080/00273171.2019.1598837

Cuijpers, P., Berking, M., Andersson, G., Quigley, L., Kleiboer, A., & Dobson, K. S. (2013). A meta-analysis of cognitive-behavioural therapy for adult depression, alone and in comparison with other treatments. *Canadian Journal of Psychiatry*, *58*(7), 376–385. https://doi.org/10.1177/070674371305800702

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

Day, F. R., Ruth, K. S., Thompson, D. J., Lunetta, K. L., Pervjakova, N., Chasman, D. I., … Murray, A. (2015). Large-scale genomic analyses link reproductive aging to

hypothalamic signaling, breast cancer susceptibility and BRCA1-mediated DNA repair. *Nature Genetics*, *47*(11), 1294–1303. https://doi.org/10.1038/ng.3412

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

de Jong, S. (1993). SIMPLS: An alternative approach to partial least squares regression. *Chemometrics and Intelligent Laboratory Systems*, *18*(3), 251–263. https://doi.org/10.1016/0169-7439(93)85002-X

de Leeuw, J., Young, F. W., & Takane, Y. (1976). Additive structure in qualitative data: An alternating least squares method with optimal scaling features. *Psychometrika*, *41*(4), 471–503. https://doi.org/10.1007/BF02296971

De Mol, C., Giannone, D., & Reichlin, L. (2008). Forecasting using a large number of predictors: Is Bayesian shrinkage a valid alternative to principal components? *Journal of Econometrics*, *146*, 318–328. https://doi.org/10.1016/j.jeconom.2008.08.011

Deng, L., Yang, M., & Marcoulides, K. M. (2018). Structural Equation Modeling With Many Variables: A Systematic Review of Issues and Developments. *Frontiers in Psychology*, *9*, 1–14. https://doi.org/10.3389/fpsyg.2018.00580

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

Edwards, J. R., & Bagozzi, R. P. (2000). On the nature and direction of relationships between constructs and measures. *Psychological Methods*, *5*(2), 155–174. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/10937327

Efron, B. (1979). Bootstrap Methods: Another Look at the Jackknife. *The Annals of Statistics*, *7*(1), 1–26. https://doi.org/10.1214/AOS/1176344552

Efron, B. (1983). Estimating the Error Rate of a Prediction Rule: Improvement on Cross-Validation. *Journal of the American Statistical Association*, *78*(382), 316. https://doi.org/10.2307/2288636

Efron, B., & Tibshirani, R. (1986). Bootstrap Methods for Standard Errors, Confidence Intervals, and Other Measures of Statistical Accuracy. *Statistical Science*, *1*(1), 54–75. https://doi.org/10.1214/ss/1177013815

Efron, & Tibshirani. (1997). Improvements on Cross-Validation: The 632+ Bootstrap Method. *Journal of the American Statistical Association*, *92*(438), 548–560. https://doi.org/10.1080/01621459.1997.10474007

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

Enki, D. G., Trendafilov, N. T., & Jolliffe, I. T. (2013). A clustering approach to interpretable principal components. *Journal of Applied Statistics*, *40*(3), 583–599. https://doi.org/10.1080/02664763.2012.749846

Evans-Polce, R. J., Veliz, P. T., Boyd, C. J., Hughes, T. L., & McCabe, S. E. (2019). Associations between sexual orientation discrimination and substance use disorders: differences by age in US adults. *Social Psychiatry and Psychiatric Epidemiology*, 1–10. https://doi.org/10.1007/s00127-019-01694-x

Farrar, D. E., & Glauber, R. R. (1967). Multicollinearity in Regression Analysis: The Problem Revisited. *The Review of Economics and Statistics*, *49*(1), 92. https://doi.org/10.2307/1937887

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

Gauchi, J. P., & Chagnon, P. (2001). Comparison of selection methods of explanatory variables in PLS regression with application to manufacturing process data. In *Chemometrics and Intelligent Laboratory Systems* (Vol. 58, pp. 171–193). Elsevier. https://doi.org/10.1016/S0169-7439(01)00158-7

Geisser, S. (1975). The Predictive Sample Reuse Method with Applications. *Journal of the American Statistical Association*, *70*(350), 328. https://doi.org/10.2307/2285815

Geladi, P., & Kowalski, B. R. (1986). Partial least-squares regression: a tutorial. *Analytica Chimica Acta*, *185*(C), 1–17. https://doi.org/10.1016/0003-2670(86)80028-9

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

Gusev, A., Ko, A., Shi, H., Bhatia, G., Chung, W., Penninx, B. W. J. H., … Pasaniuc, B. (2016). Integrative approaches for large-scale transcriptome-wide association studies. *Nature Genetics*, *48*(3), 245–252. https://doi.org/10.1038/ng.3506

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

Hittner, J. B., Penmetsa, N., Bianculli, V., & Swickert, R. (2020). Personality and substance use correlates of e-cigarette use in college students. *Personality and Individual Differences*, *152*, 109605. https://doi.org/10.1016/j.paid.2019.109605

Hotelling, H. (1957). The Relations of the Newer Multivariate Statistical Methods to Factor Analysis. *British Journal of Statistical Psychology*, *10*(2), 69–79. https://doi.org/10.1111/j.2044-8317.1957.tb00179.x

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. (2006a). 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

Hu, M.-C., Davies, M., & Kandel, D. B. (2006b). 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

Hwang, H., & Takane, Y. (2014). *Generalized structured component analysis: A component-based approach to structural equation modeling* (1st ed.). New York: Chapman and Hall/CRC. https://doi.org/10.1201/b17872

Jackson, D. L. (2001). Sample size and number of parameter estimates in maximum likelihood confirmatory factor analysis: A Monte Carlo investigation. *Structural Equation Modeling*, *8*(2), 205–223. https://doi.org/10.1207/S15328007SEM0802_3

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

Jolliffe, I. T. (1982). A Note on the Use of Principal Components in Regression. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, *31*(3), 300–303. https://doi.org/10.2307/2348005

Jöreskog, K. G. (1970). A general method for analysis of covariance structures. *Biometrika*, *57*(2), 239–251. https://doi.org/10.1093/biomet/57.2.239

Jöreskog, K. G. (1973). A generating method for estimating a linear structural equation system. In A. S. Goldberger & O. D. Duncan (Eds.), *Structural Equation Models in the Social Sciences* (pp. 85–112). Seminar Press. https://doi.org/10.1002/j.2333-8504.1970.tb00783.x

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

Kendall, M. G. (1957). *A Course in Multivariate Analysis*. London: Charles Griffen & Co. Retrieved from https://www.amazon.com/Course-Multivariate-Analysis-M-Kandall/dp/B001PKBO90

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

Kozberg, M. G., Chen, B. R., DeLeo, S. E., Bouchard, M. B., & Hillman, E. M. C. (2013). Resolving the transition from negative to positive blood oxygen level-dependent responses in the developing brain. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(11), 4380–4385. https://doi.org/10.1073/pnas.1212785110

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

Liang, K. Y., & Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, *73*(1), 13–22. https://doi.org/10.1093/biomet/73.1.13

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

MacCallum, R. C., & Browne, M. W. (1993). The use of causal indicators in covariance structure models: some practical issues. *Psychological Bulletin*, *114*(3), 533–541. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/8272469

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

Maitra, S., & Models, J. Y. (2008). Principle component analysis and partial least squares: Two dimension reduction techniques for regression. *Applying Multivariate Statistical Models*, *79*, 79–90. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.473.4340&rep=rep1&type=pdf#page=81

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.

McDonald, R. P. (1996). Path analysis with composite Variables. *Multivariate Behavioral Research*, *31*(2), 239–270. https://doi.org/10.1207/s15327906mbr3102_5

Mehmood, T., & Ahmed, B. (2016). The diversity in the applications of partial least squares: an overview. *Journal of Chemometrics*, *30*(1), 4–17. https://doi.org/10.1002/cem.2762

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

Moustafa, A. A., Diallo, T. M. O., Amoroso, N., Zaki, N., Hassan, M., & Alashwal, H. (2018). Applying Big Data Methods to Understanding Human Behavior and Health. *Frontiers in Computational Neuroscience*, *12*(84), 1–4. https://doi.org/10.3389/fncom.2018.00084

Na, K. S. (2019). Prediction of future cognitive impairment among the community elderly: A machine-learning based approach. *Scientific Reports*, *9*(1), 1–9. https://doi.org/10.1038/s41598-019-39478-7

National Institute on Drug Abuse. (2019). Drugged driving. *National Institutes of Health; U.S. Department of Health and Human Services.*, pp. 1–5. Retrieved from https://www.drugabuse.gov/publications/drugfacts/drugged-driving

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

Pan, W. (2001). Akaike's information criterion in generalized estimating equations. *Biometrics*, *57*(1), 120–125. https://doi.org/10.1111/j.0006-341X.2001.00120.x

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

Seibold, H., Zeileis, A., & Hothorn, T. (2016b). 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

Shmueli, G. (2010). To Explain or to Predict? *Statistical Science*, *25*(3), 289–310. https://doi.org/10.1214/10-STS330

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*(11), 1359–1366. https://doi.org/10.1177/0956797611417632

Smith, G., & Campbell, F. (1980). A critique of some ridge regression methods. *Journal of the American Statistical Association*, *75*(369), 74–81. https://doi.org/10.1080/01621459.1980.10477428

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

Smith, & Sasaki. (1979). Decreasing multicollinearity: A method for models with multiplicative functions. *Sociological Methods & Research*, *8*(1), 35–56. https://doi.org/10.1177/004912417900800102

Spano, V. R., Mandell, D. M., Poublanc, J., Sam, K., Battisti-Charbonney, A., Pucci, O., … Mikulis, D. J. (2013). CO2 Blood Oxygen Level–dependent MR Mapping of Cerebrovascular Reserve in a Clinical Population: Safety, Tolerability, and Technical Feasibility. *Radiology*, *266*(2), 592–598. https://doi.org/10.1148/radiol.12112795

Stone, M. (1974). Cross-Validatory Choice and Assessment of Statistical Predictions. *Journal of the Royal Statistical Society*, *36*(2), 111–147. https://doi.org/10.2307/2984809

Strobl, C., Kopf, J., & Zeileis, A. (2015a). 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., Kopf, J., & Zeileis, A. (2015b). 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

Tian, T. S., Wilcox, R. R., & James, G. M. (2010). Data reduction in classification: A simulated annealing based projection method. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, *3*(5), 319–331. https://doi.org/10.1002/sam.10087

U.S. Department of Health and Human Services. (2004). The Health Consequences of Smoking: A Report of the Surgeon General. *National Library of Medicine*, *2012*, 51576–51576. https://doi.org/10.1002/yd.20075

United States Department of Health and Human Services. Substance Abuse and Mental Health Services Administration. Center for Behavioral Health Statistics and Quality. (2013). National Survey on Drug Use and Health Database. Inter-university Consortium for Political and Social Research (ICPSR) [distributor]. https://doi.org/10.3886/ICPSR35509.v1

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

Wold, H. (1966). Estimation of principal components and related models by iterative least squares. In P. R. Krishnajah (Ed.), *Multivariate analysis* (pp. 391–420). NewYork: Academic Press.

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

Zhang, W., Du, T., & Wang, J. (2016). Deep Learning over Multi-field Categorical Data. In *Advances in Information Retrieval. ECIR 2016. Lecture Notes in Computer Science* (Ferro N. et al., Vol. 9626, pp. 45–57). Springer, Cham. https://doi.org/10.1007/978-3-319-30671-1_4

Appendix A. Parameter Estimation in PLSR

As discussed earlier, NIPALS and SIMPLS algorithms are the most commonly used algorithms for PLSR. For simplicity, the NIPALS algorithm is presented here.

1. Background

The main idea behind PLSR is to calculate the principal components of the **X** and the **Y** matrix separately (external correlation) and to develop a regression model between the scores of the principal components (inner correlation). That is, PLSR aims to obtain the decompositions of both **X** and **Y**, i.e., $\mathbf{X} = \mathbf{TP}'$ and $\mathbf{Y} = \mathbf{UQ}'$, as in PCA, and then subsequently perform regression between **T** and **U** ($\mathbf{U} = \mathbf{TB}$), where **P** is the principal components of **X** ($\mathbf{T} = \mathbf{XP}$) and **Q** is the principal components of **Y** ($\mathbf{U} = \mathbf{YQ}$). Note that **P** is the $k \times k$ orthogonal matrix obtained as $\mathbf{T} = \mathbf{XP}$ such that the columns of **T**, $\mathbf{t}_1, \ldots \mathbf{t}_k$, are uncorrelated and arranged in order of decreasing variance. **P** is often called the loading matrix and **T** is called the score matrix in the PLS literature.

If we do the above decompositions of **X** and **Y** separately using the NIPALS algorithm, each update rule is:

Decomposition of **X**	Decomposition of **Y**
X: column centered and normalized	**Y**: column centered and normalized
t: initialized with random values	**u**: initialized with random values
Loop	Loop
$\mathbf{p} = \mathbf{X}'\mathbf{t} / \|\|\mathbf{X}'\mathbf{t}\|\|$	$\mathbf{q} = \mathbf{Y}'\mathbf{u} / \|\|\mathbf{Y}'\mathbf{u}\|\|$
$\mathbf{t} = \mathbf{Xp}$	$\mathbf{u} = \mathbf{Yq}$
Until **t** stop changing	Until **u** stop changing

This results in the first principal components **p** and **q**, as well as their corresponding score vectors **t** and **u**. To find the subsequent components and score vectors, repeat the same steps

after partialing out the effect of **t** and **u** from **X** and **Y**: i.e., **X** = **X** - **tp**′ and **Y** = **Y** - **uq**′. After *l* such steps, we obtain two *N*×*l* matrices **T** and **U** for *l* < *k*, i.e., the matrices in a reduced dimension, subsequently **P** and **Q**.

2. PLSR Algorithm

As noted earlier, PLSR seeks to find the decompositions or the linear combinations of both **X** and **Y** in such a way that the covariance between the obtained linear combinations, i.e., **t**′**u**, is maximum. One intuitive way to achieve this is to exchange **t** and **u** in the update rules for **p** and **q** described above and combined the two update rules in a single loop. This results in the following procedure:

- Initially, **X** and **Y** are column-centered and normalized.
- Before starting the iteration process, **u** is initialized with random values.
- Loop

  p = **X**′**u** / ||**X**′**u**||

  t = **Xp**

  q = **Y**′**t** / ||**Y**′**t**||

  u = **Yq**

  Until **t** stop changing

  (The vectors **t**, **u**, **p**, and **q** are then stored in the corresponding matrices)
- This finds the first set of PLS components and loadings. For subsequent components and vectors, set **X** = **X** - **tp**′ and **Y** = **Y** - **uq**′, then repeat the same steps.
- After *l* such steps, we obtain **T**, **U**, **P**, and **Q**.

Then, to get a regression model relating **Y** and **X**, we first fit **B** for **U** = **TB**, i.e., **B** = **T**′**U**, subsequently, **Y** = **UQ**′ = **TBQ**′ = **XPBQ**′.

Appendix B. Estimation and Inference in Parametric ERA

We express (2.2.3) in matrix notation as

$$\varphi = (\boldsymbol{z} - \mathbf{X}\mathbf{W}\boldsymbol{b})'\boldsymbol{\Omega}(\boldsymbol{z} - \mathbf{X}\mathbf{W}\boldsymbol{b}) = (\boldsymbol{z} - \mathbf{F}\boldsymbol{b})'\boldsymbol{\Omega}(\boldsymbol{z} - \mathbf{F}\boldsymbol{b}) \tag{B1}$$

with respect to **W** and $\boldsymbol{b}$, subject to diag($\mathbf{F}'\mathbf{F}$) = $N\mathbf{I}$, where $\boldsymbol{z}$ is an N by 1 vector of adjusted response variable values z_i, **X** is an N by P matrix of predictors, **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 **F** is an N by K matrix of component scores.

To estimate ERA parameters, we aim to minimize (B1) by an iterative method in which each iteration involves the following steps:

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

$$\begin{aligned}
\varphi_{(\mathbf{W})} &= (\boldsymbol{z} - \mathbf{X}\mathbf{W}\boldsymbol{b})'\boldsymbol{\Omega}(\boldsymbol{z} - \mathbf{X}\mathbf{W}\boldsymbol{b}) \\
&= [\mathrm{vec}(\boldsymbol{z} - \mathbf{X}\mathbf{W}\boldsymbol{b})]'\boldsymbol{\Omega}[\mathrm{vec}(\boldsymbol{z} - \mathbf{X}\mathbf{W}\boldsymbol{b})] \\
&= [\boldsymbol{z} - (\boldsymbol{b}' \otimes \mathbf{X})\mathrm{vec}(\mathbf{W})]'\boldsymbol{\Omega}[\boldsymbol{z} - (\boldsymbol{b}' \otimes \mathbf{X})\mathrm{vec}(\mathbf{W})] \\
&= (\boldsymbol{z} - \mathbf{U}\boldsymbol{w}^*)'\boldsymbol{\Omega}(\boldsymbol{z} - \mathbf{U}\boldsymbol{w}^*)
\end{aligned} \tag{B2}$$

where $\otimes$ indicates the Kronecker product, vec(**W**) indicates the vec operator that creates the column vector of **W** obtained by stacking the columns of **W**, **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(**W**), and $\boldsymbol{w}^*$ denotes the P by 1 vector of the nonzero elements in vec(**W**). Then, the estimates of $\boldsymbol{w}^*$ are obtained by

$$\hat{\boldsymbol{w}}^* = (\mathbf{U}'\boldsymbol{\Omega}\mathbf{U})^{-1}\mathbf{U}'\boldsymbol{\Omega}\boldsymbol{z}\,. \tag{B3}$$

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

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

$$\begin{aligned}\varphi_{(\mathbf{b})} &= (\boldsymbol{z} - \mathbf{X}\mathbf{W}\boldsymbol{b})'\boldsymbol{\Omega}(\boldsymbol{z} - \mathbf{X}\mathbf{W}\boldsymbol{b}) \\ &= (\boldsymbol{z} - \mathbf{F}\boldsymbol{b})'\boldsymbol{\Omega}(\boldsymbol{z} - \mathbf{F}\boldsymbol{b})\end{aligned} \tag{B4}$$

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

$$\hat{\boldsymbol{b}} = (\mathbf{F}'\boldsymbol{\Omega}\mathbf{F})^{-1}\mathbf{F}'\boldsymbol{\Omega}\boldsymbol{z}\,. \tag{B5}$$

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 $\mathbf{W}$ and $\boldsymbol{b}$ 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., 2015b; 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

$$-\mathrm{H}(\boldsymbol{\theta}) = -\frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'} = -\begin{pmatrix} \dfrac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \boldsymbol{w}^* \partial \boldsymbol{w}^{*\prime}} & \dfrac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \boldsymbol{w}^* \partial \boldsymbol{b}'} \\ \dfrac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \boldsymbol{b} \partial \boldsymbol{w}^{*\prime}} & \dfrac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \boldsymbol{b} \partial \boldsymbol{b}'} \end{pmatrix}. \tag{B6}$$

The diagonal terms in (B6) can be obtained by fixing $\boldsymbol{W}^*$ and $\boldsymbol{b}$, respectively:

$$-\frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \boldsymbol{w}^* \partial \boldsymbol{w}^{*\prime}} = -\mathbf{U}'\boldsymbol{\Omega}\mathbf{U} \tag{B7}$$

and

$$-\frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \boldsymbol{b} \partial \boldsymbol{b}'} = -\mathbf{F}'\boldsymbol{\Omega}\mathbf{F}\,. \tag{B8}$$

The off-diagonal terms in (B6) can be obtained using the profile likelihoods (Richards, 1961)

$$-\frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \boldsymbol{b} \partial \boldsymbol{w}^{*\prime}} = -\frac{\partial \boldsymbol{w}^{*\prime}}{\partial \boldsymbol{b}}\left(-\frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \boldsymbol{w}^{*} \partial \boldsymbol{w}^{*\prime}}\right). \tag{B9}$$

To compute $-\frac{\partial \boldsymbol{w}^{*\prime}}{\partial \boldsymbol{b}}$ in (B9), let $\boldsymbol{\delta}_j$ denote a *K* by 1 vector of 0 except having 1 in the *j*th element (*j* = 1, …, *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(**W**). Then, $-\frac{\partial \boldsymbol{w}^{*\prime}}{\partial \boldsymbol{b}}$ is calculated by

$$-\frac{\partial \boldsymbol{w}^{*\prime}}{\partial \boldsymbol{b}} = (\mathbf{U}'\boldsymbol{\Omega}\mathbf{U})^{-1}\left[\Lambda'\boldsymbol{\Omega}\boldsymbol{z} - \Lambda'\boldsymbol{\Omega}\mathbf{U}\left((\mathbf{U}'\boldsymbol{\Omega}\mathbf{U})^{-1}\mathbf{U}'\boldsymbol{\Omega}\boldsymbol{z}\right)\right] \quad (j = 1, \ldots, K). \tag{B10}$$

Appendix C. Parameter Estimation in GEE-ERA

As mentioned in Chapter 1.4, the parameter estimation algorithm for GEE-ERA was proposed and briefly described in Lee et al. (2019). This section reiterates the algorithm in full details, using the notation of Method 3.2 for consistency.

To minimize (3.2.3), GEE-ERA uses an iterative algorithm that repeats the following steps until the changes in the estimated parameter values between previous and current iterations are below a pre-determined threshold, e.g., 10^{-5} :

Step 1. Update **B** for fixed **W** and $R_i(\boldsymbol{\alpha})$. This is equivalent to minimizing the following criterion with respect to **B**,

$$
\begin{aligned}
\varphi_{(\mathbf{B})} &= \sum_{i=1}^{N}[(\tilde{z}_i - \mathbf{B}'\tilde{f}_i)'\Sigma_i^{-1}(\tilde{z}_i - \mathbf{B}'\tilde{f}_i)] + \lambda_{\mathbf{B}}\mathrm{tr}(\mathbf{B}'\mathbf{B}) \\
&= \mathrm{tr}\left(\sum_{i=1}^{N}[(\tilde{z}_i - \mathbf{B}'\tilde{f}_i)'\Sigma_i^{-1}(\tilde{z}_i - \mathbf{B}'\tilde{f}_i)] + \lambda_{\mathbf{B}}\mathrm{tr}(\mathbf{B}'\mathbf{B})\right) \\
&= \sum_{i=1}^{N}\mathrm{tr}\left([(\tilde{z}_i - \mathbf{B}'\tilde{f}_i)'\Sigma_i^{-1}(\tilde{z}_i - \mathbf{B}'\tilde{f}_i)]\right) + \lambda_{\mathbf{B}}\mathrm{tr}(\mathbf{B}'\mathbf{B}) \\
&= \sum_{i=1}^{N}\mathrm{tr}\left([(\tilde{z}_i - \mathbf{B}'\tilde{f}_i)'\Sigma_i^{-1}(\tilde{z}_i - \mathbf{B}'\tilde{f}_i)]\cdot\mathbf{I}\right) + \lambda_{\mathbf{B}}\mathrm{tr}(\mathbf{B}'\mathbf{B}) \\
&= \sum_{i=1}^{N}\mathrm{vec}(\tilde{z}_i - \mathbf{B}'\tilde{f}_i)'\cdot(\mathbf{I}\otimes\Sigma_i^{-1})\cdot\mathrm{vec}(\tilde{z}_i - \mathbf{B}'\tilde{f}_i) + \lambda_{\mathbf{B}}\mathrm{tr}(\mathbf{B}'\mathbf{B}) \\
&= \sum_{i=1}^{N}\mathrm{vec}(\tilde{z}_i - \mathbf{B}'\tilde{f}_i)'\ \Sigma_i^{-1}\ \mathrm{vec}(\tilde{z}_i - \mathbf{B}'\tilde{f}_i) + \lambda_{\mathbf{B}}\mathrm{tr}(\mathbf{B}'\mathbf{B})\ ,
\end{aligned}
\tag{C1}
$$

where $\tilde{f}_i$ is a K by 1 vector of the component scores of the ith respondent, **I** is the identity matrix, tr(**A**) indicates the trace of a square matrix A, $\otimes$ indicates the Kronecker product, vec(**A**) indicates the vec operator that creates the column vector of **A** obtained by stacking the columns of **A**. Let vec(**B'**) = ***b***, then

$$
\begin{aligned}
\varphi_{(\mathbf{B})} &= \sum_{i=1}^{N} \left(\mathrm{vec}(\tilde{z}_i) - \mathrm{vec}(\mathbf{B}'\tilde{f}_i) \right)' \ \Sigma_i^{-1} \ \left(\mathrm{vec}(\tilde{z}_i) - \mathrm{vec}(\mathbf{B}'\tilde{f}_i) \right) + \lambda_{\mathbf{B}} \mathrm{tr}(\mathbf{B}'\mathbf{B}) \\
&= \sum_{i=1}^{N} (\tilde{z}_i - (\tilde{f}_i' \otimes \mathbf{I}) \cdot \boldsymbol{b})' \ \Sigma_i^{-1} \ (\tilde{z}_i - (\tilde{f}_i' \otimes \mathbf{I}) \cdot \boldsymbol{b}) + \lambda_{\mathbf{B}} \boldsymbol{b}'\boldsymbol{b} \\
&= \sum_{i=1}^{N} (\tilde{z}_i - \mathbf{Q}_i \boldsymbol{b})' \ \Sigma_i^{-1} \ (\tilde{z}_i - \mathbf{Q}_i \boldsymbol{b}) + \lambda_{\mathbf{B}} \boldsymbol{b}'\boldsymbol{b} ,
\end{aligned}
\tag{C2}
$$

where $\mathbf{Q}_i = (\tilde{f}_i' \otimes \mathbf{I})$. Then, the estimates of $\boldsymbol{b}$ are obtained by

$$
\hat{\boldsymbol{b}} = \left[\sum_{i=1}^{N} \mathbf{Q}_i' \Sigma_i^{-1} \mathbf{Q}_i \ + \ \lambda_{\mathbf{B}} \cdot \mathbf{I} \right]^{-1} \left[\sum_{i=1}^{N} \mathbf{Q}_i' \Sigma_i^{-1} \ \tilde{z}_i \right], \tag{C3}
$$

and subsequently, the nonzero elements in **B** are replaced with the corresponding values in $\hat{\boldsymbol{b}}$.

<u>Step 2.</u> Update **W** for fixed **B** and $R_i(\boldsymbol{\alpha})$. This is equivalent to minimizing the following criterion with respect to **W**,

$$
\begin{aligned}
\varphi_{(\mathbf{W})} &= \sum_{i=1}^{N} [(\tilde{z}_i - \mathbf{B}'\mathbf{W}'\tilde{x}_i)' \Sigma_i^{-1} (\tilde{z}_i - \mathbf{B}'\mathbf{W}'\tilde{x}_i)] + \lambda_{\mathbf{W}} \mathrm{tr}(\mathbf{W}'\mathbf{W}) \\
&= \sum_{i=1}^{N} \mathrm{tr}\left([(\tilde{z}_i - \mathbf{B}'\mathbf{W}'\tilde{x}_i)' \Sigma_i^{-1} (\tilde{z}_i - \mathbf{B}'\mathbf{W}'\tilde{x}_i)] \right) + \lambda_{\mathbf{W}} \mathrm{tr}(\mathbf{W}'\mathbf{W}) \\
&= \sum_{i=1}^{N} \mathrm{vec}(\tilde{z}_i - \mathbf{B}'\mathbf{W}'\tilde{x}_i)' \ \Sigma_i^{-1} \ \mathrm{vec}(\tilde{z}_i - \mathbf{B}'\mathbf{W}'\tilde{x}_i) + \lambda_{\mathbf{W}} \mathrm{tr}(\mathbf{W}'\mathbf{W}) \\
&= \sum_{i=1}^{N} (\tilde{z}_i - (\tilde{x}_i' \otimes \mathbf{B}') \cdot \boldsymbol{w})' \ \Sigma_i^{-1} \ (\tilde{z}_i - (\tilde{x}_i' \otimes \mathbf{B}') \cdot \boldsymbol{w}) + \lambda_{\mathbf{W}} \boldsymbol{w}'\boldsymbol{w} \\
&= \sum_{i=1}^{N} (\tilde{z}_i - \mathbf{M}_i \boldsymbol{w})' \ \Sigma_i^{-1} \ (\tilde{z}_i - \mathbf{M}_i \boldsymbol{w}) + \lambda_{\mathbf{W}} \boldsymbol{w}'\boldsymbol{w} ,
\end{aligned}
\tag{C4}
$$

where $\mathrm{vec}(\mathbf{W}') = \boldsymbol{w}$ and $\mathbf{M}_i = (\tilde{x}_i' \otimes \mathbf{B}')$. The estimates of $\boldsymbol{w}$ are obtained by

$$
\hat{\boldsymbol{w}} = \left[\sum_{i=1}^{N} \mathbf{M}_i' \Sigma_i^{-1} \mathbf{M}_i \ + \ \lambda_{\mathbf{W}} \cdot \mathbf{I} \right]^{-1} \left[\sum_{i=1}^{N} \mathbf{M}_i' \Sigma_i^{-1} \ \tilde{z}_i \right], \tag{C5}
$$

and the nonzero elements in **W** are replaced with the corresponding values in $\hat{\boldsymbol{w}}$.

<u>Step3.</u> Update $R_i(\boldsymbol{\alpha})$ for fixed **B** and **W**. More specifically, the correlation parameters in $\boldsymbol{\alpha}$ are estimated from the current Pearson residuals defined by

$$\hat{r}_{iq} = (y_{iq} - \hat{\mu}_{iq}) / \text{var}(\hat{\mu}_{iq})^{1/2}, \tag{C6}$$

where $\hat{\mu}_{iq}$ depends on the current values for **B** and **W**. As mentioned in the Method section, the estimator for **α** depends upon the choice of $R_i(\boldsymbol{\alpha})$. See Liang & Zeger (1986, pp. 17–18) for the specific estimators. Finally, the scale parameter ϕ is estimated by

$$\hat{\phi} = \sum_{i=1}^{N} \sum_{q=1}^{Q} \hat{r}_{iq}^{\,2} / (NQ - (K + P)). \tag{C7}$$