Contemporary Issues in Exploratory Data Mining in the Behavioral Sciences

Edited by John J. McArdle and Gilbert Ritschard

Routledge
Taylor & Francis Group
NEW YORK AND LONDON

First published 2014
By Routledge
711 Third Avenue, New York, NY 10017

Simultaneously published in the UK
By Routledge
27 Church Road, Hove, East Sussex BN3 2FA

Routledge is an imprint of the Taylor & Francis Group, an informa business

Library of Congress Cataloging in Publication Data
Contemporary issues in exploratory data mining in the behavioral sciences / edited by John J. McArdle, University of Southern California and Gilbert Ritschard, University of Geneva.
pages cm. – (Quantitative methodology series)
1. Social sciences–Statistical methods. 2. Social sciences–Research–Data processing. 3. Data mining–Social aspects.
I. McArdle, John J. II. Ritschard, Gilbert, 1950–
HA29.C746 2014
006.3'12–dc23 2013006082

ISBN: 978-0-415-81706-6 (hbk)
ISBN: 978-0-415-81709-7 (pbk)
ISBN: 978-0-203-40302-0 (ebk)

Typeset in Bembo
by Cenveo Publisher Services

SFI Certified Sourcing
www.sfiprogram.org
SFI-00453

Printed and bound in the United States of America
by Edwards Brothers, Inc.

4 Exploratory Data Mining with Structural Equation Model Trees

Andreas M. Brandmaier, Timo von Oertzen, John J. McArdle, and Ulman Lindenberger

Introduction

Structural Equation Model Trees (SEM Trees) combine Structural Equation Models (SEM) and decision trees. SEM Trees are tree structures that partition a dataset recursively into subsets with significantly different sets of parameter estimates. The method allows the detection of heterogeneity observed in covariates and thereby offers the possibility to automatically discover non-linear influences of covariates on model parameters in a hierarchical fashion. The methodology allows an exploratory approach to SEM by providing a data-driven but hypothesis-constrained exploration of the model space. We summarize the methodology, show applications on empirical data, and discuss Hybrid SEM Trees, an extension of SEM Trees that allows the finding of subgroups that differ with respect to model parameters and model specification.

In this chapter, we present an overview and selected applications of a multivariate statistical framework, Structural Equation Model Trees (SEM Trees; Brandmaier, von Oertzen, McArdle, & Lindenberger, 2013), that combines benefits from confirmatory and exploratory approaches to data analysis. SEM Trees allow a data-driven refinement of models reflecting prior hypotheses about the data. Confirmatory aspects are provided by using Structural Equation Modeling as the framework, and exploratory aspects arise from the incorporation of decision trees, also known as *classification and regression trees* or *recursive partitioning*. This combined approach yields trees representing a recursive partitioning of a dataset into subgroups maximally differing with respect to their model-predicted distributions. SEM Trees allow an exploratory approach to finding variables that influence the model parameters for any model that can be described as a linear combination of observed and latent variables. This class of models includes, for instance, regression models (McArdle & Epstein, 1987), factor analytic models (Jöreskog, 1969), autoregressive models (Jöreskog, 1979; McArdle & Aber, 1990), latent growth curve models (McArdle & Epstein, 1987), latent difference score models (McArdle &

Hamagami, 2001), or latent differential equation models (Boker, Neale, & Rausch, 2004).

A typical workflow for empirical research in the behavioral sciences can be described by the following steps. First, hypotheses about the population are derived as tentative explanations of observed phenomena. Then, a study is designed and conducted to collect a dataset with variables representing concrete observations of the phenomena. Finally, each hypothesis is formalized in a model and inference-statistical methods are used to gauge the evidence of the data for or against the hypotheses. This is often called the *confirmatory* approach of data analysis. Unfortunately, it is often found that models describe the data inadequately. Consequently, researchers move to an exploratory phase, in which they adapt hypotheses and models, in order to find a better representation of the observed phenomena, for instance, by adding variables or removing variables from their models. As an alternative approach to improving the model as a description of the complete dataset, a second approach can be followed: The dataset is partitioned into groups that differ with respect to the parameter estimates of the model. This multi-group approach assumes that the model is a valid description of the phenomena, however, it does not require the sample to be homogeneous with respect to the parameters of the model. SEM Trees realize this approach by recursively partitioning a dataset into subgroups that maximally differ in the model-predicted distributions.

In this chapter, we summarize the algorithm for inducing SEM Trees and highlight details such as the estimation of parameters in the models, the evaluation of split candidates, and the incorporation of measurement invariance in the SEM Tree framework. We draw attention to the dual motivation of the candidate selection procedure from a statistical and an information-theoretic point of view. As a methodological innovation, we discuss Hybrid SEM Trees that allow the specification of a set of different SEMs representing competing hypotheses about the data. Hybrid SEM Trees allow the retrieval of partitions of the dataset that are best described by different models and, ultimately, by different hypotheses. In order to illustrate the utility of SEM Trees, we conclude with an application of regular SEM Trees and Hybrid SEM Trees on empirical datasets.

All reported analyses are based on the freely available *semtree* package (Brandmaier, 2012b) for the statistical computing language R (Ihaka & Gentleman, 1996). The package is based on OpenMx (Boker et al., 2011) for defining and estimating SEMs.

Decision Trees

Decision trees are classifiers that discriminate between states of a response variable based on a hierarchy of decisions on a set of covariates. Put in a statistical context, a decision tree represents partitions of the covariate space that are associated with significant differences in the response variable.

The earliest representative, called the *Automatic Interaction Detector* (AID), was devised by Sonquist and Morgan (1964). The paradigm gained popularity through the seminal works of Breiman, Friedman, Olshen, and Stone (1984) and Quinlan (1986). Many aspects of decision trees have been developed since then. We restrict ourselves to mentioning only a small selection of the various approaches available today: ID3 (Quinlan, 1986), C4.5 (Quinlan, 1993), CART (Breiman et al., 1984), and QUEST (Loh & Shih, 1997). A more recent development are model-based trees. These trees maximize differences of outcome variables with respect to a hypothesized model, e.g., logistic regression trees (Chan & Loh, 2004), multivariate adaptive regression splines (MARS; Friedman, 1991), or the comprehensive model-based partitioning framework by Zeileis, Hothorn, and Hornik (2008). Decision trees are increasingly used in the context of ensemble learning. In this paradigm, multiple decision trees, typically based on resampled subsets of the original dataset, are aggregated into a decision forest, e.g., random forests (Breiman, 2001) or conditional random forests (Strobl, Boulesteix, Zeileis, & Hothorn, 2007), thereby trading increased computation time and decreased interpretability for increased stability and predictive accuracy of the results.

In the machine learning community, the popularity of tree methods gradually diminished in favor of more recent learning machine methods that allowed learning about more complex decision boundaries. Nevertheless, decision trees can be helpful in the process of knowledge discovery and scientific theory building due to their clear advantage in depicting the predictive structure visually. Figure 4.1 illustrates decision boundaries of a two-class problem obtained from a logistic regression as a representative of linear discriminating models and the corresponding boundaries produced by a decision tree. From a decision-tree perspective, a linear model can be thought of as an *oblique tree* restricted to a height of one. Oblique trees (Murthy, Kasif, & Salzberg, 1994), sometimes referred to as *multivariate trees* (Brodley & Utgoff, 1995), allow decisions to be represented as linear combinations of covariates in inner nodes of the tree. While oblique trees allow a larger number of possible splits, traditional decision tree approaches with axis-parallel splits have the advantage of allowing a straightforward interpretation of the decision tree as simple decision rules in natural language, specifically in rule sets describing conditions on covariates, e.g., "IF a participant is younger than 25 AND works out regularly, THEN she/he has a low risk of heart disease."

The result of a recursive partitioning algorithm is typically visualized in a dendrogram (cf. Figure 4.1). A dendrogram is a pictorial description of the hierarchical differences in the model-implied predictions and the predictors that determine these differences. Ovals represent decision nodes. The label of a decision node contains the name of the covariate that is subject to partitioning. If the partitioning is based on a statistical test, a corresponding p-value or a test statistic is shown in the label. Each oval has two or more

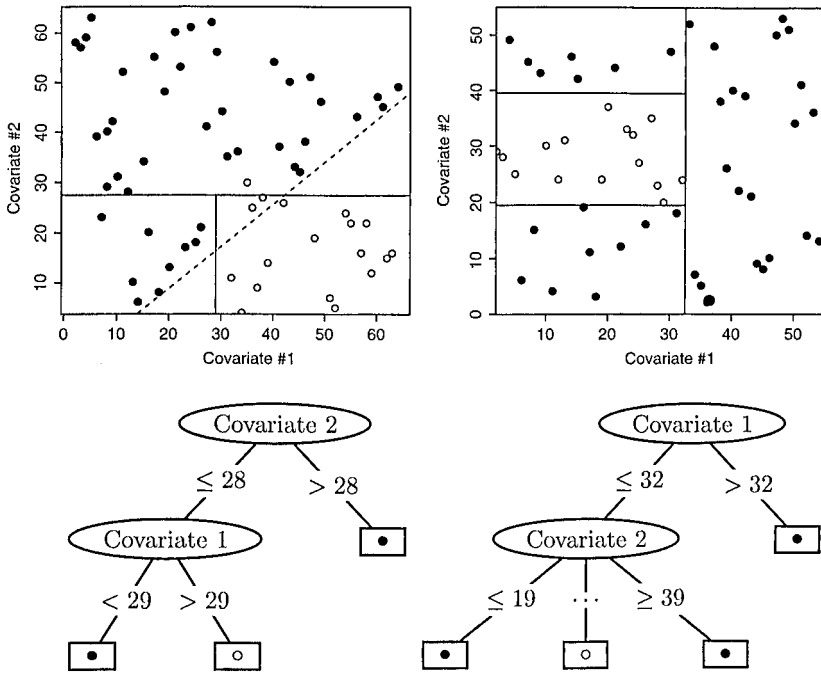

Figure 4.1 The top row shows two-dimensional decision boundaries on tv hypothetical datasets with two classes depicted as empty and solid do On the left, the solid lines depict the axis-parallel decision boundaries the tree below that discriminates the two classes. The dotted line depi the decision boundary of a logistic regression discriminating betwe both classes. On the right, the solid lines represent axis-parallel decisi boundaries of the decision tree below. In the example on the rig adequate discrimination is not possible using a linear model.

outgoing edges that represent the partitioning of the dataset into subse corresponding to whether the covariate value of the oval node match the condition that is depicted on the respective edge. Leaf nodes of t trees are depicted as rectangles that contain information about the predict outcome; in the case of SEM Trees, they contain parameter estimates f the chosen SEM.

Structural Equation Model Trees

Hyafil and Rivest (1976) showed that finding an optimal tree is NP-ha (which can only be solved in non-deterministic polynomial time), where optimality is defined as a minimization of the expected number of decisio and the underlying problem is computationally demanding to solve. Th motivates the widespread application of heuristics for the induction

decision trees. When applying decision trees, the general idea is to choose the covariate from a set of candidates that divides the dataset into groups that maximally increase the predictability of the outcome variable. This process is recursively applied to each resulting partition of the dataset as long as meaningful covariates are found. The "goodness-of-split" can be formalized in various ways, for example, based on information-theoretic or statistical tests.

SEM Trees are based on the idea that datasets can not only be partitioned into subgroups that are homogeneous with respect to a single outcome variable, but into subgroups that are homogeneous with respect to the parameters of an SEM. In the behavioral and social sciences, Structural Equation Models (SEMs) have become widely accepted as a statistical tool for modeling the relations between latent and observed variables. SEMs are based on an isomorphism between (a) a set of linear equations for observed and latent variables and distributional assumptions about these variables and (b) a graphical representation of these equations and assumptions.

SEM Trees recover decision boundaries in the covariate space dividing the dataset into multiple groups that are each represented by a different parameter set for an SEM. If Figure 4.1 represented covariate space boundaries of an SEM Tree, one could imagine the solid and empty dots representing participants with two different associated parameter sets.

The algorithm for the induction of an SEM Tree is geared to the traditional decision tree algorithms. Thus, it is a greedy, top-down, recursive partitioning procedure that chooses the locally best split of the covariate space. In each recursion step, it chooses the covariate that is maximally informative about the model-predicted distribution and, according to this choice, permanently splits the dataset. Inputs of the algorithm are: (1) an SEM that formalizes the researchers' hypotheses about the data, which is also referred to as *template model*; (2) an empirical dataset that is modeled by the SEM; and (3) a set of covariates whose influence on the SEM is to be explored. In a first step, parameters of the template model are estimated from the complete dataset. This parameter set is associated with the root of a decision tree. For each covariate, the dataset is temporarily partitioned into subgroups according to the values of the covariate. Then, parameters are estimated for each partitioned dataset. We refer to the model that estimates parameters on the unsplit dataset as the *pre-split model*. The set of models resulting from splitting the dataset into subgroups can be seen as a single multiple-group model, which we refer to as the *post-split model*. If the parameter estimates are obtained with maximum likelihood estimation, the pre-split model and the post-split model are algebraically nested models and their log-likelihood ratio is asymptotically χ^2-distributed under the null hypothesis that the covariate is uninformative about the model-predicted distribution (Brandmaier, 2012a). At each level of the tree, the covariate with the maximum log-likelihood ratio is chosen. This selection procedure is recursively continued in each resulting partition. The known

Algorithm 4.1 The elementary algorithm for the induction of SEM Trees with discrete covariates and multi-way splits. Typically, a preprocessing step reduces all covariates to sets of binary covariates yielding a binary tree.

InduceSEMTree (*Dataset*, *Covariates*, *Model*)

1. Create a new *node*
2. Estimate free parameters $\hat{\theta}$ from *Dataset*
3. If *Covariates* is empty, return *node*
4. For each covariate $C_i \in$ *Covariates*,
   a. For each value v_{ij} of the covariate C_i,
      i. Estimate $\hat{\theta}_{v_i}$ from the subset $D_{v_{ij}} \subset D$ for which covariate C_i has value v_{ij}
   b. Calculate the log-likelihood ratio statistic Λ_{C_i} for covariate C_i as:

$$\Lambda_{C_i} = -2\mathcal{LL}\left(M\left(\hat{\theta}\right)|D\right) + \sum_j 2\mathcal{LL}\left(M\left(\hat{\theta}_{v_{ij}}\right)|D_{v_{ij}}\right)$$

5. Find best covariate candidate $b = \arg\max_i \Lambda_{C_i}$
6. If Λ_{C_b} is above the critical value,
   a. For each v_{bj} in C_b create a child node by recursively calling InduceSEMTree ($D_{v_{bj}}$,*Covariates*\C_b,*Model*) and create an edge between *node* and the new child node with label v_{bj}
7. Otherwise return *node*

---

distributional properties of the estimator under the null hypothesis allow the usage of a hypothesis-testing framework to determine when to stop splitting the dataset. If there is no covariate having a significantly large test statistic such that the null hypothesis can be rejected, the induction algorithm terminates. The elementary algorithm for the induction of an SEM Tree with discrete covariates is shown in Algorithm 4.1. This algorithm allows multi-way splits according to the levels of discrete covariates. In the following, a generalization to continuous covariates is described.

Multi-valued Attributes

Drawing upon ideas of Breiman et al. (1984) and Loh and Shih (1997), SEM Trees perform dichotomous splits in the covariates leading to binary tree representations. Generally, trees with multiway splits can be represented as binary trees. However, Kim and Loh (2001) illustrate an example that yields different results depending on whether a multiway and a binary tree are used.

To allow continuous, and multi-valued ordinal and discrete covariates, all multi-valued covariates are transformed to sets of dichotomous covariates. The conversion depends on the type of variable. Ordinal variables having values from an ordered set are transformed into a set of dichotomous variables, in which each variable represents a "smaller or equal" relation on one of the possible split points. Continuous variables with values from an ordered set imply cutpoints in the center between pairs of sorted

observed values. Let N be the number of observations in a dataset. With the proposed procedure, ordinal and continuous variables can imply a maximum of $N-1$ dichotomous covariates. Categorical covariates imply dichotomous covariates representing splits corresponding to partitions in all possible pairs of non-empty subsets yielding a maximum of $2^{N-1}-1$ dichotomous covariates.

Model Estimation

Parameters in models of an SEM Tree are estimated using maximum likelihood estimation, which is a common technique in SEM (e.g., von Oertzen, Ghisletta, & Lindenberger, 2009). Let X be a dataset. Let M be a template SEM that encodes researchers' prior hypotheses about the data. Under the assumption of independence of the observations in the sample, the likelihood of the model given the dataset with a total of N observations is the product of the likelihoods of observing the model given individual observations. Let Σ be the model-implied covariance matrix and let μ be the model-implied mean vector. Under the assumption that the variables are normally distributed, the likelihood of the model given a single datum x is defined by the multivariate Gaussian distribution:

$$\mathcal{L}(\mu, \Sigma|x) = ((2\pi)^p |\Sigma|)^{-\frac{1}{2}} \exp\left[-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right] \quad (1)$$

Given a set of observations, let m be the sample mean vector and S be the sample covariance matrix. Furthermore, let θ be a vector parametrizing μ and Σ. The simplified log-likelihood function derived from Equation 1 is:

$$\begin{aligned} &-2\mathcal{LL}(\mu, \Sigma, \theta|m, S) \\ &\quad = N\left[\text{const} + \log|\Sigma| + tr\left(\Sigma^{-1}S\right) + (m-\mu)^T \Sigma^{-1}(m-\mu)\right] \end{aligned} \quad (2)$$

The maximum likelihood estimate of the parameters $\hat{\theta}$ given the data m, S, and the model μ, Σ is found by minimizing $-2\mathcal{LL}$:

$$\hat{\theta} = \arg\min_{\theta} -2\mathcal{LL}(\mu, \Sigma, \theta|m, S) \quad (3)$$

Because it is difficult to find a closed-form solution for general models, numeric procedures are employed to find the maximum of the likelihood function or the minimum of the negative function. There are a variety

of methods for the numerical solution of the problem. Most solutions revolve around the *Newton method* in order to find the minimum of the function numerically. The gradient, i.e., the partial derivatives of the likelihood function with respect to the parameters, describes the rate of increase or decrease of the likelihood function depending on an infinitesimally small change of the free parameters. *Gradient descent* methods propose iteratively calculating the gradient, and climbing or descending the likelihood function until they arrive at a maximum or minimum. An improvement of this approach also considers the matrix of the partial second derivatives, the *Hessian matrix*. This comprises the class of Newton methods. If the Hessian matrix is iteratively approximated rather than fully approximated at each iteration, this is called a *quasi-Newton method*. An important and widely used representative of the latter method is the *Broyden, Fletcher, Goldfarb, and Shanno method* (BFGS), which was independently suggested by each of the four authors. A comprehensive overview of optimization algorithms is given by Fletcher (1994). OpenMx (Boker et al., 2011), on which the *semtree* package is based, uses a general-purpose optimization scheme that involves numerical estimation of the gradient and the Hessian matrix. Von Oertzen et al. (2009) report that a dampened Newton method, which fully calculates the gradient and the Hessian at each step and adapts the step width by a line search, works well in practice.

Split Candidate Evaluation

A recursive tree-inducing algorithm proceeds by selecting the best candidate at each step of the recursion. There are natural stopping criteria for a recursive tree-inducing algorithm, including the following: (1) there are no remaining covariates to split the dataset; (2) the number of observations in a leaf node is below a certain threshold; (3) a pre-determined height of the tree has been reached; and (4) the best split candidate is not good enough. The fourth criterion is introduced in order to avoid overfitting, i.e., to avoid choosing an apparently adequate split candidate although its aptitude is only due to random fluctuation in the sample. In SEM Trees, split candidate selection can be based on the log-likelihood ratio that allows a statistical test to determine when splitting the dataset and growing the tree should be stopped.

Given a θ-parametrized model $M(\theta)$ and a dataset D, let $\hat{\theta}$ be a parameter vector that minimizes the *negative two log-likelihood* of seeing the model given the data, i.e., the maximum-likelihood estimate of M given D. Let $D_{v_{ij}}$ be the partitions of a dataset with respect to the $j = 1, \ldots, k$ values of the ith covariate in the dataset and let $\hat{\theta}_{v_{ij}}$ be the parameter vector that minimizes $-2\mathcal{LL}(M(\hat{\theta}_{v_{ij}})|D_{v_{ij}})$. For covariate i, the likelihood ratio of the

pre-split model (left summand) and the post-split model of the $j = 1, \ldots, k$ resulting partitions (right summand) is:

$$\Lambda_i = -2\mathcal{LL}\left(M(\hat{\theta})|D\right) + \sum_{j=1}^{k} 2\mathcal{LL}\left(M(\hat{\theta}_{v_{ij}})|D_{v_{ij}}\right) \tag{4}$$

Following Wilks's (1938) theorem, Λ_i is χ^2-distributed with $(k-1)m$ degrees of freedom with m being the number of free parameters in the template SEM. Given the distributional properties of Λ_i and a chosen α level, a critical value c can be calculated such that $Pr(\Lambda > c) = \alpha$. Only split candidates for which $\Lambda_i > c$ holds are considered potential split candidates. The best split is chosen by selecting the covariate with the maximum log-likelihood ratio, that is, the covariate with the largest evidence against the pre-split model.

It has been pointed out that any tree-structured algorithm carrying out an exhaustive search in the split attribute space suffers from a multiple comparison problem (Jensen & Cohen, 2000) that can lead to overfitting (i.e., an over-representation of apparent structure in the sample that is merely due to sampling fluctuations). Solutions for tackling this problem include correcting the critical value under the assumption of independence of covariates, known as Bonferroni correction, or using cross-validation to obtain estimates of the expected log-likelihood ratio. The first approach corrects the sampling distribution under independence assumptions at low computational costs but is known to be overly conservative. The latter approach yields a score that can be treated as an individual score instead of a maximum score but requires additional computations.

Several authors (Dobra & Gehrke, 2001; Jensen & Cohen, 2000; Loh & Shih, 1997; Shih, 2004; Zeileis et al., 2008) have cautioned about the problem of variable selection bias in the context of decision trees. By definition, split candidate selection procedures that preferably select certain types of variables over others under the null hypothesis that all variables are uninformative, suffer from variable selection bias. A typical observation is that categorical variables with a larger number of categories are more likely to be chosen under the null hypothesis than those with less. An unbiased variable selection algorithm is expected to have no preference for any variable under the null hypothesis. Brandmaier et al. (2013) showed that using the likelihood-ratio-based split selection procedure as described above suffers from selection bias. This bias can be reduced to a neglible amount if variable cutpoints and the selection between variables are estimated in a two-step procedure (Kim & Loh, 2001). As an alternative approach, Zeileis et al. (2008) suggested a unified framework for unbiased model-based recursive partitioning that is based on tests for parameter instability.

Missing Values

Under the assumption that variable values are *missing at random* (Rubin, 1976), SEM Trees can handle missing values in the observed variables of the model as well as in the covariates. If values are missing in the covariates, the likelihood calculation is performed with Full Information Maximum Likelihood (FIML; Finkbeiner, 1979), which is equivalent to Equation 2 under the assumption of no missingness:

$$-2\mathcal{LL}(\mu, \Sigma | x_1, \ldots, x_N) = N \cdot p \cdot \ln(2\pi) + \sum_{i=1}^{N} \left[\ln|\Sigma_i| + (x_i - \mu_i)^T \Sigma_i^{-1} (x_i - \mu_i) \right] \quad (5)$$

where Σ_i is the model-implied covariance matrix with rows and columns deleted according to the pattern of missingness in the ith observation x_i, and μ_i the model-implied mean vector with elements deleted according to the respective pattern of missingness.

Missing values in covariates can be handled by removing the respective missing rows in the dataset during the evaluation of a split candidate, effectively modifying only Equation 4 based on the patterns of missingness in the dataset. Others, e.g., Hastie, Tibshirani, and Friedman (2001), employ a surrogate approach, which is based on finding a surrogate covariate that most closely describes the same partition of the dataset as the variable with the missing value.

Measurement Invariance

A fundamental issue in psychometrics is measurement invariance. A measurement is invariant if "under different conditions of observing and studying phenomena, measurement operations yield measures of the same attribute" (Horn & McArdle, 1992, p. 117). Measurement invariance is traditionally examined through a sequence of hypothesis tests. Typically, a set of statistical tests are carried out to determine what level of measurement invariance is tenable. These tests are administered in the following order: (1) *configural invariance* (also *configuration invariance* or *pattern invariance*) requires the invariance of the pattern of zero and non-zero factor loadings across groups; (2) *metric invariance*, *weak invariance*, or *factor pattern invariance* requires the invariance of the values of factor loadings across groups; (3) *strong factorial invariance* or *scalar invariance* requires intercepts of all indicators and all factor loadings to be equal across groups; (4) *strict invariance* establishes the additional restriction that the residual error variances are equal across groups, in order to allow the interpretation of standardized coefficients across groups.

SEM Trees incorporate the concept of measurement invariance if desired by the researcher. Models with measurement invariance are integrated in

the process of split candidate evaluation in the following way. Valid split candidates must fulfill the user-specified level of measurement invariance *and* their log-likelihood ratio must exceed the chosen critical value. By construction, a measurement-invariant post-split model of configural, metric, strong, or strict invariance is algebraically nested in a non-invariant post-split model. Furthermore, the invariant model is nested in the pre-split model (Brandmaier, 2012a). Consolidating these observations, one can determine valid split candidates by first performing a set of likelihood-ratio tests in order to assure measurement invariance for the split candidates. With the set of candidates for which measurement invariance could not be rejected, the normal procedure of split candidate evaluation (see above) is then performed.

Information-Theoretic Interpretation

From a machine learning perspective, split candidate selection procedures for traditional decision trees with a categorical outcome variable maximize the predictability of the outcome variable conditional on the knowledge of the state of the selected predictors. This is often formalized as follows. Let $H(X)$ be the Shannon entropy of a random variable X, let $x_1, \ldots, x_N$ be observed outcomes of X, and let N_y be the number of observations for which $Y = y$. The information gain about X when knowing the state of Y is:

$$Gain(x_1, \ldots, x_N, Y) = H(X) - \sum_{y \in Values(Y)} \frac{N_y}{N} H(x_1, \ldots, x_n | Y = y) \quad (6)$$

Let M be a model. The duality between the Gaussian log-likelihood and the entropy of the model-predicted distribution can be formulated as:

$$\lim_{N \to \infty} \frac{1}{N} \sum_{i=1}^{N} \mathcal{LL}(x_i | M) = -H(X|M) \quad (7)$$

Estimating the parameters of an SEM by maximizing transformations of the multivariate Gaussian fit function seen in Equation 2 minimizes the entropy of the model-implied distribution. The likelihood-ratio test statistic that is used to determine the significance of a split candidate is the difference between the log-likelihood of the parent model and the sum of the log-likelihoods of the potential child models. It can be shown that the likelihood-ratio test statistic from Equation 4 is proportional to the information gain shown in Equation 6 (see Brandmaier, 2012a, for a detailed proof). This important relation motivates the variable selection approach in

SEM Trees from both a statistical and an information-theoretic perspective. At each level, an SEM Tree chooses the covariate that maximizes the likelihood ratio statistic of the pre-split and the post-split model. This is equivalent to choosing the covariate that maximizes the information gain about the model-predicted distribution. The expected information gain is the mutual information. By using cross-validation as a variable selection procedure, the expected likelihood-ratio is estimated and, following the same reasoning as above, the expected information gain is maximized at each level of the decision tree.

Hybrid Trees

Thus far, SEM Trees were defined such that their goal is a recursive partitioning of a dataset into subsets that maximize the difference of the model-implied distributions. As an extension to that, Brandmaier et al. (2013) suggested that it can make sense to allow not only a single template model but a set of competing template models. This implies that leaf nodes in the tree are associated with subgroups represented by both different SEM and individual parameter estimates. We call these types of trees *Hybrid SEM Trees*. Hybrid SEM Trees can answer research questions that involve a choice between multiple candidate models for the representation of a dataset. Instead of providing a single choice of a "best" model for the complete dataset, Hybrid SEM Trees retrieve different models for different subgroups in the sample.

Suppose a group of researchers is interested in cognitive development over the life span. They suspect that some participants have a linear increase in cognitive abilities that saturates at some point, whereas a competing hypothesis assumes a simple linear change process without saturation, and a third hypothesis proposes a drastic exponential decline in cognitive change that was observed for very old participants. Rather than fitting a single model to the complete dataset, which might imply an unfavorable trade-off between the different observed phenomena of change, a Hybrid SEM Tree can recover subgroups that are best described by individual models.

We will outline the underlying model selection problem in Hybrid SEM Trees in more detail. Comparing a set of different models for subgroups renders the likelihood ratio test for model selection inappropriate because, in general, the set of competing models is not algebraically nested. However, the evaluation of split candidates by an estimate of their predictive performance is still feasible. As suggested before in the context of SEM Trees with a single template model, we employ k-fold cross-validation to obtain an estimate of the expected likelihood ratio of a pair of candidate models. The cross-validated test statistic $\hat{CV}$ is obtained by averaging the k test statistics of a k-fold cross-validation. Let M_1 and M_2 be two competing models and let x be an observation, and n be the sample size.

We obtain:

$$\lim_{n\to\infty} \hat{CV} = E\left[\log P(x|M_1) - \log P(x|M_2)\right] = E\left[\log \frac{P(x|M_1)}{P(x|M_2)}\right]$$

where $E()$ is the expectation. For model selection, we require a procedure to determine the evidence for the model. Using Bayes' formula, we can rewrite the test statistic as:

$$E\left[\log \frac{P(x|M_1)}{P(x|M_2)}\right] = E\left[\log \frac{P(M_1|x)\,P(M_2)\,P(x)}{P(M_2|x)\,P(M_1)\,P(x)}\right]$$

Under the assumption that, a priori, all candidate models are equally likely, $P(M_1) = P(M_2)$, we obtain a model selection procedure selecting covariates that maximize the expected posterior probability of the post-split model:

$$\lim_{n\to\infty} \hat{CV}_{P(M_1)=P(M_2)} = E\left[\log \frac{P(M_1|x)}{P(M_2|x)}\right]$$

Under the assumptions of non-identical priors for the models, a correction term can be added to the cross-validation statistic that represents the models' prior ratio. Effectively this is the sum of the estimated log-likelihood ratio and the difference of the log-priors.

Hybrid SEM Trees allow model selection between a set of competing SEMs representing competing hypotheses about the data, whereas recursive partitioning elicits a hierarchy of covariates associated with differences in the dataset. In the process of induction of a Hybrid SEM Tree, covariates are selected that not only maximize differences with respect to a single model-implied distribution but also select between parametrized distributions that are the best representation for the observed phenomena.

With hybrid trees, we distinguish between heterogeneity with respect to models and heterogeneity with respect to parameters. Common SEM Trees find heterogeneity with respect to parameters assuming that the model holds for all subgroups of the dataset. Additionally, Hybrid SEM Trees are able to choose between competing models. When applying Hybrid SEM Trees with a single template model, this is equivalent to applying a common SEM Tree with cross-validation for variable selection.

Case Studies

In this section, we present applications of SEM Trees to selected datasets, including a univariate SEM Tree, a regression SEM Tree, and a longitudinal SEM Tree.

Univariate SEM Tree and Regression SEM Tree

In order to illustrate the usage of univariate SEM Trees, we show an analysis based on a freely available dataset from the *psych* package (Revelle, 2011) for R. The dataset includes scores of 700 participants on the Scholastic Aptitude Test (SAT) and the American College Test (ACT) that were collected as part of the Synthetic Aperture Personality Assessment (Revelle, Wilt, & Rosenthal, 2009). Three additional covariates are included for each participant: sex (male/female), education (high school, high school graduate, college, college graduate, or graduate degree), and age ($\mu = 25.6 \pm 9.5$).

Suppose a researcher is interested in performance differences on the SAT verbal (SATV) scale. Therefore, the researcher sets up a fully saturated SEM that measures the mean of the SATV score as μ_{SATV} and the variance of the score as σ^2_{SATV}.

For the analysis, the SAT scores were standardized by subtracting the mean and dividing by the standard deviation. The SEM Tree that was generated from these data with Bonferroni-corrected p values and an alpha level of .01, is shown in Figure 4.2. The tree chooses education as the first partition in a group depending on whether participants graduated from college or not. For those who graduated from college, a second split according to their age is the optimal partition. In order to estimate the difference between the two age groups for graduates, we first calculated the pooled standard deviation of both age groups $\sigma^2 = 0.70$ and obtained an effect size of $d = 0.59$ for a split at an age of 28. Naturally, the question arises whether this particular age reflects a fixed change point of the investigated phenomenon. In general, researchers should be careful

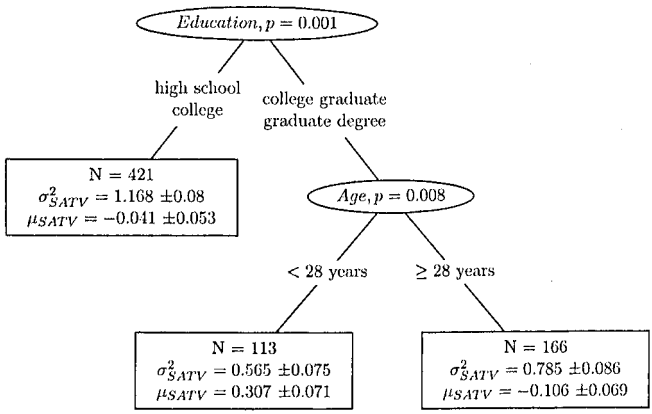

Figure 4.2 SEM Tree for the SAT dataset. Variables "education" and "age" are chosen to split the dataset into subgroups that describe differences in the SAT scores. Parameter estimates are given as point estimates with standard error.

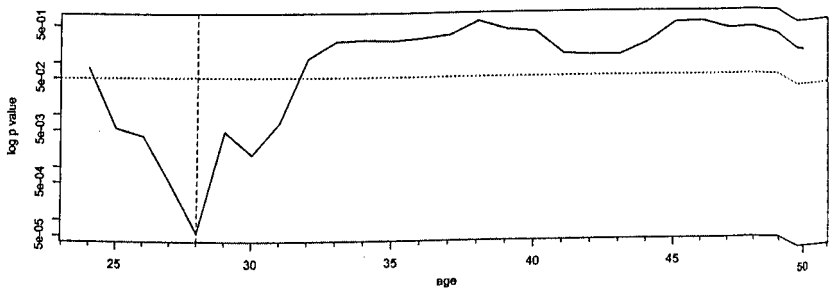

Figure 4.3 The plot shows p values for the splits according to the variable "age" conditioned on people with an education equivalent to college degree or higher. This corresponds to the right subtree of the SEM Tree in Figure 4.2. The p values are shown on a logarithmic scale. The horizontal, dotted line indicates a significance level of 5%. Values below the threshold are potential split candidates. The minimum value is attained when a split at an age of 28 years is chosen (marked by the vertical, dashed line). The set of significant split points of the age variable ranges from ages 25 to 31.

when reifying continuous splits. We advise further inspection of the tree and the dataset in order to determine an appropriate range of a continuous covariate that supports significant splits. In this example, we inspected the p values of all possible age-related splits. The p values were already corrected for multiple testing. Figure 4.3 shows the log p values versus possible age-related splits. A clear minimum is visible at an age of 28. However, the range from 25 to 31 is below a significant threshold of .05. We conclude that this age range should be interpreted as a fuzzy partitioning or fuzzy decision rule for describing the age-related subgroups instead of cresting at 28. This type of ex-post analysis can be carried out for any covariate with more than two ordered levels.

In the following, we use the same dataset to investigate a regression model with SEM Trees. Beyond the verbal score of the SAT, the dataset contains self-reports of the quantitative score on the SAT. Again, both scores were normalized to obtain zero mean and unit variance. A regression model between the two scores can reveal insights about how strongly these two scores are correlated or how well one score can predict the other score assuming a linear relationship between the two variables. The model contains five parameters. A plot of the model's path diagram is shown in Figure 4.4.

The correlation between the verbal and the quantitative score was $r = 0.64$, $p < 2.2 \times 10^{-16}$. Based on this measure, the proportion of variance shared between the two variables in the sample is about 41% under the assumption that the population is homogeneous with respect to the correlation of these abilities. Applying an SEM Tree to this model allows the

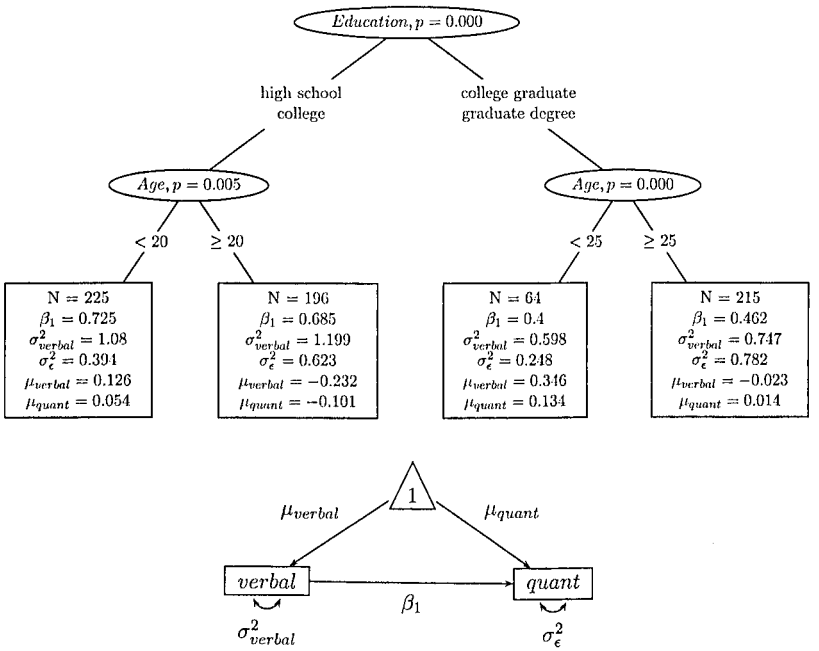

Figure 4.4 SEM Tree (upper panel) for a regression model (lower panel) of the quantitative and the verbal SAT score.

detection of subgroups that differ in the strength of the linear relationship between the two scores. Differences in the parameter β_1 between groups indicate a different strength of the linear relation between the two variables. Differences between the parameter μ_{verbal} or μ_{quant} indicate differences in the expected values of the scores between subgroups. Differences in the residual error term, σ^2_ϵ, between groups indicate differences in the model fitness, particularly, differences in the amount of variance unexplained by the linear model. The SEM Tree was induced using a significance threshold of .01. The resulting tree is shown in Figure 4.4. The subgroups that are implied by the first-level split of the SEM Tree are shown as scatterplots in Figure 4.5, in which solid and dotted regression lines indicate the linear relation between the two variables as retrieved by the SEM Tree. The variable "education" that was chosen as a first split explains differences in the linear relation between the quantitative and the verbal SAT scores. The better-educated subgroup shows a smaller linear relation between the scores. Splits with respect to age are found in both education-related subgroups. Generally, we can observe a decrease in the average performance on both scales in each of the subgroups with higher age. The difference in the cutpoint of the continuous variable "age" in the left and right subtree

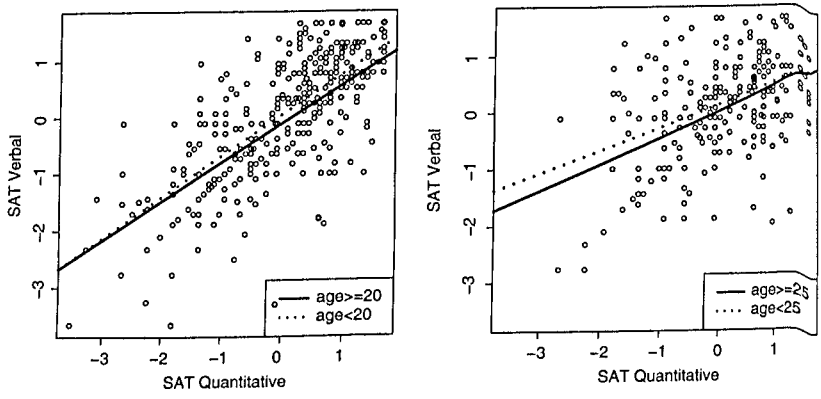

Figure 4.5 Regression plots of the quantitative and verbal SAT scores for the subgroups implied by the first split of the SEM Tree. The split variable was education and the two partitions were high school, high school graduation or college (left plot) and college graduation or graduate degree (right plot). The regression lines depict the linear relation between both scores for the second-level splits of the SEM Tree, both with respect to the covariate age.

might indicate a lag in the underlying age-related change process that is explained by "education."

Without SEM Trees, a researcher might have stopped when finding a highly significant correlation between test scores. SEM Trees proved useful here because they recovered subsets that differ with respect to the model parameters. In the tree, we found hints that the covariate "age" predicts lower average scores on both tests, whereas "education" predicts a difference in the shared variance of the scores.

Longitudinal SEM Tree

Brandmaier et al. (2013) showed illustrations of SEM Trees based on a factor model and a latent growth curve model. In the following, we further demonstrate how SEM Trees can be employed to explore structure in the data. We present an application of SEM Trees to data from the Berlin Aging Study (BASE; cf. Baltes & Mayer, 1999; Lindenberger, Smith, Mayer, & Baltes, 2010; Lövdén, Ghisletta, & Lindenberger, 2004). BASE is a multidisciplinary study of aging with extensive measurements from psychology, sociology and social policy, internal medicine and geriatrics, and psychiatry. The first wave of measurements started in 1990. The sample consisted of 516 participants who were recruited from the city registry of Berlin, Germany. The initial sample was stratified by sex and by age with a mean age of 84.9 years and an age range of 69.7–103.1 years. It was followed up longitudinally in seven further waves until 2009. For our

analysis, we relied on data from the first six waves, of which the last was completed in 2005.

For illustration purposes, we focus the analysis on the digit letter test, a measure of perceptual speed as a marker of cognitive functioning. Participants performed the task on 11 occasions in six waves spread over fifteen years. Ghisletta and Lindenberger (2004) have modeled the digit letter task with latent growth curve models before. They report that the digit letter task displayed both reliable fixed and random linear time effects but no statistically significant quadratic effects of time for the mean and variance of a quadratic slope factor. Therefore, a linear latent growth curve model was used to model changes in the performance of individuals over time. A latent intercept variable I with mean μ_I and variance σ_I^2 models the baseline performance on the task, and a latent linear slope variable S with mean μ_S and variance σ_S^2 models the increase or decrease of performance over time in the study. A correlation between the latent intercept and the latent slope was estimated as σ_{IS}^2. The exact individual time points of measurement were available for each participant and for each occasion of measurement. This allowed accurate modeling of the cognitive change process with individual but fixed slope loadings for each participant by employing these individual time points as definition variables on the slope's loadings. Put differently, the SEM can be thought of as a multi-group model with each participant being its own group, and each group being characterized by individual but fixed slope loadings, whereas each freely estimated parameter is restricted to be equal across groups. Furthermore, the slope loadings were individually centered at the mean age of the observed time span for each participant. Also, participants' age was controlled for at the latent level by adding age as a covariate with loadings onto intercept and slope. The corresponding latent growth curve model is depicted in Figure 4.6.

Due to the age of the participants, mortality led to high attrition and 61.67% of the measurements are missing. Therefore, FIML estimation was used to estimate the parameters in the model.

Covariates included education and newspaper reading and book reading habits, of each participant's father and mother. The reading habit variables were encoded on an ordinal scale with the values "often," "sometimes," "seldom," and "never." The education index was encoded on an ordinal scale with three values: "elementary school without apprenticeship," "elementary school with apprenticeship," and "secondary school certificate."

The resulting SEM Tree is depicted in Figure 4.7. The first chosen covariate is the education variable. The decline in cognitive score is lower for the higher-educated group (first-level split, left subtree). For the lower-educated group, there is a difference based on their fathers' reading behavior. For participants whose fathers read newspapers sometimes or often, the decline in perceptual speed is comparable to the decline in the better-educated group. However, if fathers did not or seldom read newspapers,

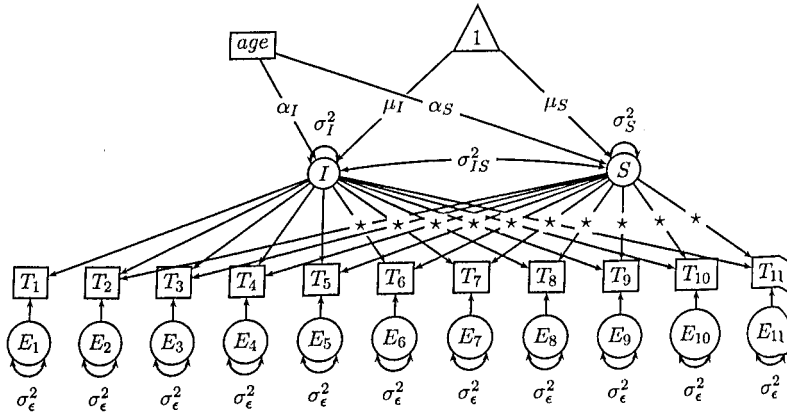

Figure 4.6 Linear latent growth curve model with individual time points for the measurement occasions. Participants were measured on 11 occasions spread over six waves. The residual error for each measurement has a variance of σ^2_ϵ. Repeated measurements of the digit letter score are represented as T_i with i being the measurement occasion. The measurement error for each observation is accordingly named E_i. The latent trajectory is modeled with an intercept $I \sim N(\mu_I, \sigma^2_I)$ and a slope $S \sim N(\mu_S, \sigma^2_S)$. The correlation between intercept and slope is modeled as σ^2_{IS}. The factor loadings of the slope are marked with stars to indicate that they are individual but fixed for each participant. The variable *age* is controlled for with loadings α_I and α_S on the latent level.

the decline is about 27% stronger than in the other group. Possibly, the fathers' reading behavior acts as a proxy for parents' education and the tree might depict an interaction of children's and fathers' educational level in predicting cognitive decline in old age.

Factor-analytic SEM Tree

For this example, we analyze a personality dataset available as "bfi" in the *psych* package (Revelle, 2011) for R. The dataset includes 25 personality self-report items taken from the International Personality Item Pool for 2,700 participants. For an illustration, we set up a single factor model as an SEM that models the personality trait "extraversion" as follows: five observed variables $X_1, X_2, \ldots, X_5$ model five items related to the factor extraversion with items including "make friends easily" or "find it difficult to approach others." Each score has an individual measurement error $E_1, E_2, \ldots, E_5$ with individual residual variances $\sigma^2_{\epsilon_1}, \sigma^2_{\epsilon_2}, \ldots, \sigma^2_{\epsilon_5}$. The means of the items were modeled as $\mu_1, \mu_2, \ldots, \mu_5$. The latent factor "*ext*"

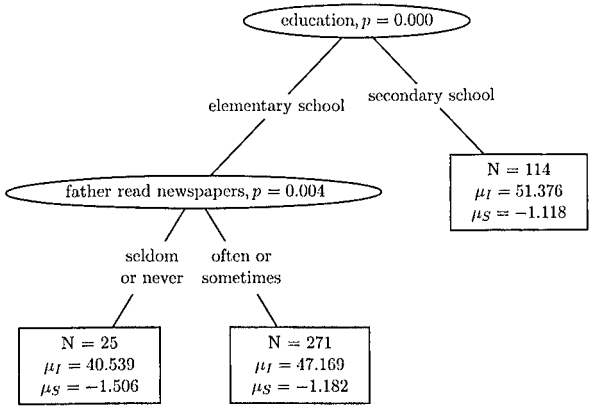

Figure 4.7 An SEM Tree based on a linear latent growth curve model. Candidate covariates included four variables of parents' reading behavior and a variable about the participants' education.

was modeled to have a mean of μ_{ext} and a variance of σ^2_{ext}. Figure 4.8 depicts the factor model.

The resulting SEM Tree was constructed with the requirement of weak invariance, that is, factor loadings were required to be equal in subgroups, and by using the Bonferonni-corrected variable selection procedure. Covariates include the variables gender, education, and age. The tree is shown in Figure 4.9. It has two levels. The first split is with respect to the age covariate, the second split is conditional on whether participants were younger than thirty years. On the second level, only the younger group is partitioned according to gender, whereas the older group is not. Comparing the estimates of the nodes, we find a difference in the means of the latent variable. Participants older than 30 years and females younger than 30 have comparable scores on the extraversion factor, both above average, whereas the young males have a below-average value of "extraversion." The interpretation of the differences can be guided by inspecting the set of estimates in the leaf nodes. For example, a comparison of the residual errors of items $(\sigma^2_{\epsilon 1}, \ldots, \sigma^2_{\epsilon 5})$ might show that variance in certain items is less well explained by the common factor in one group than in the other. In this example, this does not seem to be the case. Differences in the mean values for the individual items $(\mu_{x2}, \ldots, \mu_{x5})$ could hint at systematic differences. Note that some researchers might recommend a stricter level of measurement invariance to draw conclusions from the differences. This could be a level of measurement error that does not allow differences in the expected values of the items. Nevertheless, building a tree with weaker constraints can give insights into which covariates induce subgroups that maximally break this requirement.

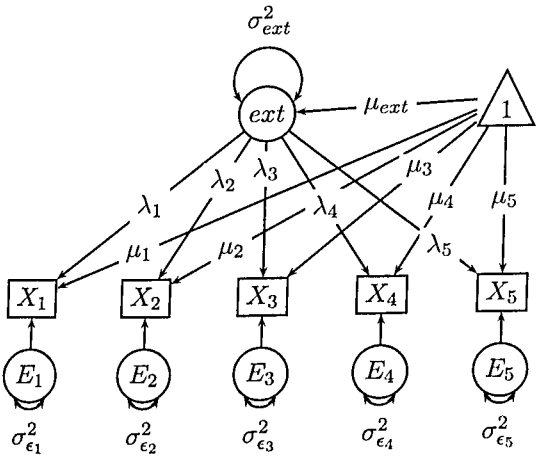

Figure 4.8 A single factor SEM for the personality factor "extraversion." The factor *ext* is measured by five items, X_1 to X_5.

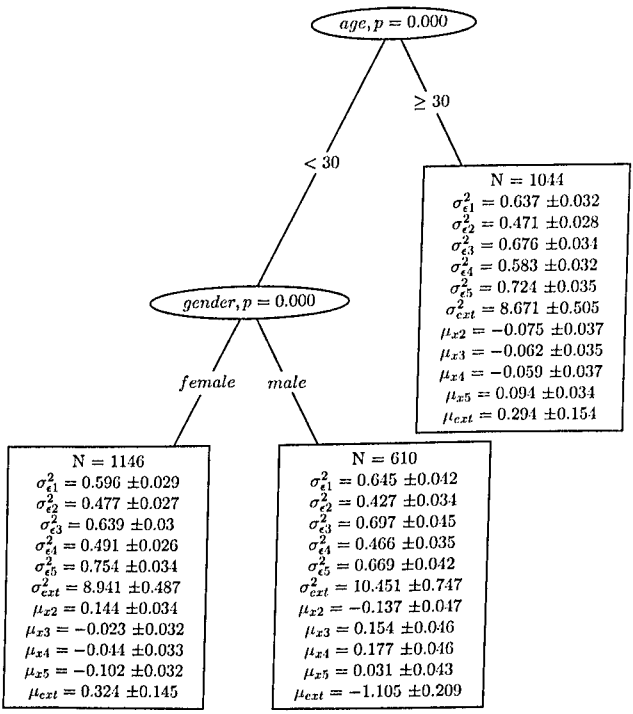

Figure 4.9 SEM Tree for an extraversion factor from a personality questionnaire.

A Hybrid SEM Tree simulation

For an illustration of a Hybrid SEM Tree, we first simulated a dataset measuring hypothetical cognitive decline in younger and older age. In this hypothetical sample, participants are described by two dichotomous covariates that carry the following meanings: (a) participants were sampled from young and older adults, and (b) participants were either part of a training program or a control group. Assume a group of researchers is interested in finding a model that describes cognitive development in their sample. Two template latent growth curve models were constructed for the Hybrid SEM Tree, one describing a model of linear change and one describing a model of quadratic change. The linear model assumes that a datum $x_{i,t}$ describing the score of individual i at time point t is an observation of the following generative model:

$$x_{i,t} = I_i + (t - t_0)\, S_i + E_i$$

with I being an intercept term that is distributed with mean μ_I and σ_I^2, S a slope term being distributed with μ_S and σ_S^2, and E being a residual error term that is distributed with zero mean and a variance σ_ϵ^2. The quadratic growth curve model is analogously created with a quadratic instead of a linear growth term:

$$x_{i,t} = I_i + (t - t_0)^2\, S_i + E_i$$

In our simulated dataset, younger adults have an approximately linear decline in their cognitive abilities that is mitigated by the cognitive training program, whereas older adults have an accelerating, quadratic decline over time. Participants were either male or female, which had no effect on cognitive development for any subgroup. The dataset was simulated with the following values. For all participants the residual error had a variance of $\sigma_\xi^2 = .01$ and the covariance between intercept and slope was set at $\sigma_{IS}^2 = 0$. The younger participants without treatment were simulated with an intercept of $\mu_I = 0$ and $\sigma_I^2 = 1$, and a slope of $\mu_I = -0.8$ and $\sigma_S^2 = .25$. The younger participants with treatment were simulated with an intercept of $\mu_I = 0$ and $\sigma_I^2 = 1$, and a slope of $\mu_I = -1.6$ and $\sigma_S^2 = .25$. Independently of the received treatment, older participants were simulated with an intercept of $\mu_I = 0$ and $\sigma_I^2 = 1$, and a quadratic slope component of $\mu_I = -0.8$ and $\sigma_S^2 = .25$.

An example SEM Tree that was obtained from randomly creating a dataset with the given values is depicted in Figure 4.10. A graphical representation of the expected growth curves of the subgroups, as they were detected by the tree, is also shown in Figure 4.10. The tree finds two subgroups with respect to age in the first split. In the young group,

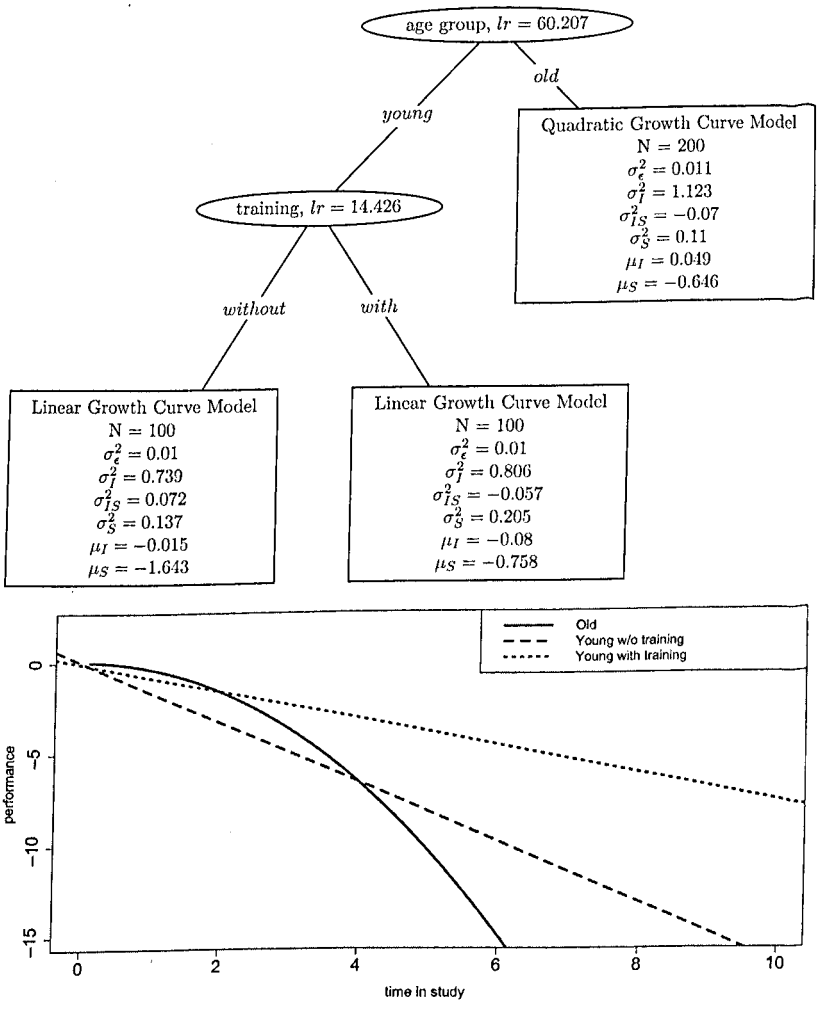

Figure 4.10 Analysis of a simulated dataset using a hybrid SEM Tree with two template models: a linear latent growth curve model and a quadratic latent growth curve model. Upper panel: the SEM Tree shows a first partition with respect to the age group. Older participants experience a quadratic decline, while younger participants are better described by a linear decline. Treatment has an influence on the slope of the change process for younger participants while no significant parameter differences for the older group were found on the second level of the tree. Lower panel: expected growth curves over time in the study for the three subgroups of the simulated dataset that were recovered by the tree.

the tree partitions the sample according to treatment. Most noteworthy, the hybrid tree correctly chose different models for the young and old subtrees. Young participants are represented with a linear model of change and old participants' accelerating decline is represented by a quadratic change. Of course, this example can also be phrased in a confirmatory setting, in which group differences with respect to the polynomial growth are expected and tested in a multi-group model. We expect Hybrid SEM Trees to work well when (1) there are a number of competing hypotheses, (2) models are complex, and (3) the number of covariates is large and their influence on the individual models and the nature of their interactions is not known (e.g., if researchers have additional sets of behavioral, cognitive, or genetic covariates).

Hybrid SEM Tree with a Developmental Latent Growth Curve Model: Wechsler Intelligence Score for Children

The data for the following illustration of an application of Hybrid SEM Trees were originally obtained by Osborne and Suddick (1972) between 1961 and 1965. These data have been analyzed in depth before (e.g., McArdle & Epstein, 1987; McArdle, 1988). An analysis with SEM Trees was performed by Brandmaier et al. (2013) using a linear latent growth curve model. We extend this analysis to a Hybrid SEM Tree.

The dataset was created from measurements for 204 children on eleven different items from the *Wechsler Intelligence Scale for Children* (WISC; Wechsler, 1949). The children were repeatedly measured on four occasions, at the ages of six, seven, nine, and eleven years. The raw scores of four "verbal" subscales and four "performance" subscales were aggregated into a composite score, which was rescaled to a range between 0 and 100. In this analysis, the covariates included the dichotomous variables "sex," the continuous variable "age," and the continuous variable "years of education" of each mother and father. The covariate "father's education" had missing values. We set up an equivalent latent growth curve model for an analysis with SEM Trees based on the description by Brandmaier et al. (2013). Then, we created modified versions of this model representing competing hypotheses about the underlying growth curve of cognitive development. All candidate models were derived from the following baseline model (see Figure 4.11), which we refer to as *BASELINE*. This model has four observed variables representing the test scores at the four occasions of measurement. Each occasion has independent errors of measurement. Five freely estimated parameters describe the distribution at the latent level: The intercept is assumed to be distributed as $I \sim \mathcal{N}(\mu_I, \sigma_I^2)$, the slope term is assumed to be distributed according to $S \sim \mathcal{N}(\mu_S, \sigma_S^2)$, and the covariance between both is modeled as σ_{IS}^2. The first slope loading was fixed at zero and the remaining slope loadings are parametrized as $\lambda_1 \lambda_2$, and λ_3. The residual variance σ_ϵ^2 is assumed to be equal for each occasion

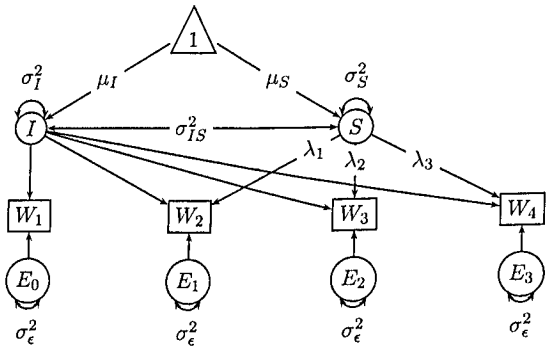

Figure 4.11 Longitudinal latent growth curve model for the WISC dataset. A composite score of WISC is measured longitudinally on four occasions represented by variables W_1, W_2, W_3, and W_4 with corresponding errors of measurement $E_{1,\ldots,4}$. Model parameters represent the latent intercept $I \sim \mathcal{N}(\mu_I, \sigma_I^2)$, the latent slope $S \sim \mathcal{N}(\mu_S, \sigma_S^2)$, the covariance between both σ_{IS}^2, the residual error terms at each occasion with variance σ_ϵ^2 and the slope loadings beyond the first occasion of measurement λ_1, λ_2, and λ_3.

of measurement. We derived four candidate models that are nested in the template model.

1 A linear latent growth curve model represents the hypothesis that cognitive development in this time period is approximately linear. In order to obtain this model from BASELINE, we set $\lambda_1 = 1$, $\lambda_2 = 3$, and $\lambda_3 = 5$. This model is referred to as *LINEAR*.
2 A hypothesis of no cognitive change is formalized in the *FLAT* model, by removing the slope component from BASELINE and thereby eliminating the parameters μ_S, σ_S^2, σ_{IS}^2, λ_1, λ_2, and λ_3.
3 We assume that cognitive improvement stops after an individual maximum level of improvement that is attained at the third occasion. This STOPPING model is achieved by setting $\lambda_1 = 1$ and adding the constraint $\lambda_2 = \lambda_3$ to BASELINE.
4 The fourth model SATURATION is a variant of the previous model that assumes a linear trajectory of cognitive development on the first three occasions and allows a different slope for the fourth occasion of measurement that can represent either a saturation of change or a boost in change. This is achieved by setting $\lambda_1 = 1$, $\lambda_2 = 3$ and freely estimating λ_3.

A Hybrid SEM Tree was induced from the dataset using the set of all four described template models, which reflect competing hypotheses about

the expected trajectories of cognitive development. The resulting SEM Tree with Bonferroni-corrected p values is shown in Figure 4.12 and the expected growth trajectories that are implied by the models associated with the four leaf nodes of the tree are plotted in Figure 4.13. The covariate "father's education" constitutes the primary partition. In each subset, "mother's education" is the second split covariate. The selection of the split point is the same as reported by Brandmaier et al. (2013).

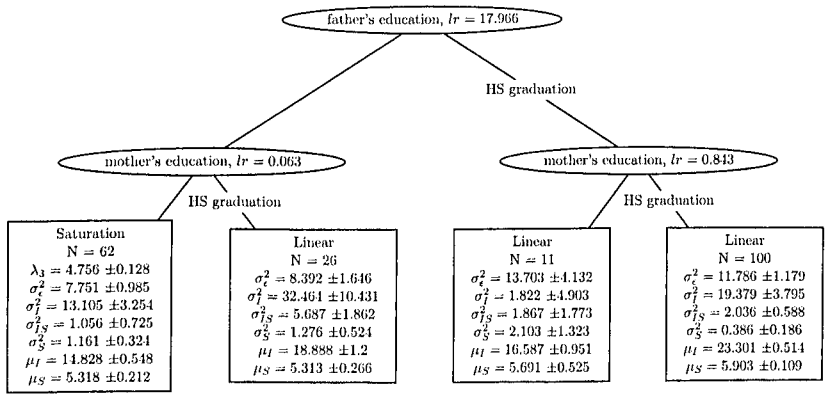

Figure 4.12 Hybrid SEM Tree on the WISC dataset. The LINEAR model representing linear change throughout all occasions of measurement is chosen for all subgroups with the exception of the subgroup in which both parents did not graduate from high school (HS; leftmost leaf). Parameter estimates are given with their standard deviations. *lr* indicates the expected log-likelihood ratio.

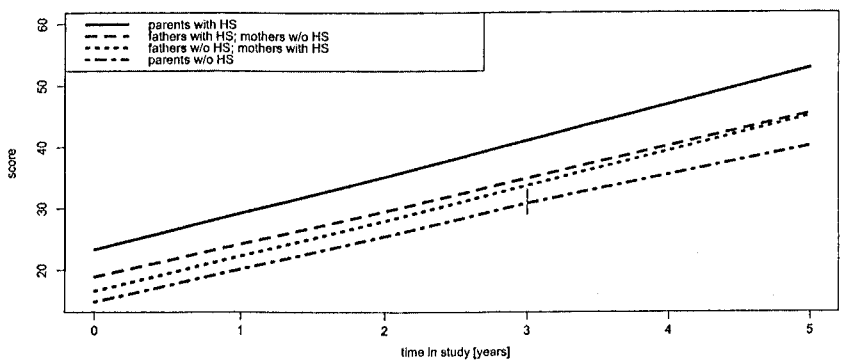

Figure 4.13 Expected growth trajectories for the subgroups retrieved by an SEM Tree with a height of two. The trajectories differ in their intercepts depending on whether parents graduated from high school (HS) or not. The linear growth of the dot-dashed trajectory exhibits a slight saturation at time point 3, as marked by a vertical line.

Among the ordinal values representing different qualities of education, the maximally informative split for both variables is related to whether parents graduated from high school or not. The difference in effect size of the slope difference between both extreme groups, the group with the higher-educated parents and the group with the less-educated parents, is quite high with an effect size of $d = .76$. However, they seem to be dominated by differences in the mean trajectories (see Figure 4.13). While previous analyses have focussed on the assumption of an approximately linear developmental change, we discovered an interesting effect with the hybrid analysis. The cognitive trajectory of children with parents that did not graduate from high school (leftmost leaf in the tree, see Figure 4.12) is better represented by the SATURATION model that allows a change of slope between the last two occasions of measurement. This hints that the cognitive development of the children from this subgroup saturates earlier than for those from the remaining groups, for which the linear model assumes a fixed $\lambda_3 = 5$, while it is estimated as $\lambda_3 = 4.76$ for the low-performing group, representing a less pronounced increase between the last two occasions. This can be observed as a slight flattening of the expected developmental trajectory of this subgroup, shown in Figure 4.13.

Conclusion

In this chapter we have reviewed the SEM Tree methodology and highlighted important methodological and algorithmic aspects. Furthermore, we have contributed an extension to the paradigm: outlining a Hybrid SEM Tree methodology that not only allows model parameters to differ across subgroups but also across model specification. We have concluded with empirical examples to demonstrate how SEM Trees can be used to find influences of covariates on models and model parameters.

We have provided several examples of how SEM Trees can be applied to empirical data. Based on SAT scores, we have shown trees with a univariate and a bivariate regression model as template models. In the former tree, the continuous covariate "age" was chosen as the split variable. We presented an analysis of the different possible age-related split points that discourages reification of the binary age-split but rather suggests reporting a range of changes. Furthermore, we have presented a longitudinal SEM Tree based on cognitive data from the BASE study and showed the influence of parents' reading behavior, presumably as a proxy of educational background, on children's cognitive development in old age. In addition, we have shown a factor-analytic SEM Tree that identifies an interaction of age and gender in the extraversion score on a personality test. We have concluded with a demonstration of Hybrid SEM Trees that allow a set of potentially non-nested template models instead of a single model. First, we illustrated the method with simulated data and proceeded to demonstrate

it on an empirical dataset of the Wechsler Intelligence Score for Children. Between a set of longitudinal SEMs that represent competing hypothesis of cognitive change, the SEM Tree selects a linear model that has been similarly described before but also hints that there is a subgroup for which a change point model could be a better representation.

In the context of longitudinal data, SEM Trees are related to longitudinal recursive partitioning, as introduced by Segal (1992), including the extension presented by Zhang and Singer (1999). Su, Wang, and Fan (2004) suggested building trees based on models estimated with maximum-likelihood procedures. An alternative framework for SEM Trees is provided in the R package *pathmox* (Sanchez & Aluja, 2012) by Sanchez (2009) that is based on partial least squares estimation of linear models. Zeileis et al. (2008) proposed a general framework for recursive partitioning based on permutation tests available in the package *party* (Hothorn, Hornik, Strobl, Zeileis, & Hothorn, 2011). Merkle and Zeileis (2011) suggested use of the R package *strucchange* (Zeileis, Leisch, Hornik, & Kleiber, 2002) to create trees based on factor models that recover subgroups that maximally break specified invariance assumptions. Further research comparing the different estimation methods and covariate selection procedures remains to be done.

Not all datasets are equally suited for the application of recursive partitioning. Due to the successive partitioning of the dataset, the resulting subsets quickly reduce in size. Generally, the larger a dataset, the better it is suited to recursive partitioning. As a rule of thumb, Hawkins (1999) has suggested that the sample size should have at least three digits in order to apply a recursive partitioning algorithm. However, the required sample size depends on many factors including number and type of free parameters, effect size, missingness, normality of the data, and choice of estimator. Also, if decision boundaries are linear but not orthogonal to the axes, or if decision boundaries are complex, decision trees tend to yield large and overly complex tree descriptions. Furthermore, the framework assumes that heterogeneity in the dataset is observed, that is, that covariates were obtained that elicit meaningful group difference with respect to the model. If unobserved heterogeneity is assumed, latent mixture models (McLachlan & Peel, 2000; Lee & Song, 2003) might provide a good starting point for further analyses. However, the clear advantage of trees is their straightforward interpretability as a "white box" model and the possibility of searching a large space of covariates and covariate interactions for influences on the model-predicted distribution.

Critics tend to allege that researchers reporting exploratory results cannibalize chance and are simply dredging data instead of carefully excavating reliable patterns. They claim that the exploitation of a large hypothesis space generates results that are merely random fluctuations and that exploration compromises their confirmatory results. Addressing the former problem, SEM Trees are equipped with procedures that support

the generalizability of findings, for example, by controlling statistical error or employing cross-validation for model and variable selection. The latter criticism is addressed in the propositions of ethical data analysis by McArdle (2010), in which he advocates performing exploratory analysis *after* the confirmatory analysis.

To conclude, SEM Trees provide a versatile exploratory data analysis tool for SEM given that a set of covariates is available whose influence on the model is as yet unclear. The method combines exploratory detection of influences of these covariates on parameter estimates for observed variables, latent variables, and their relations, and formal confirmatory mechanisms to ensure generalizability. An implementation of SEM Trees providing a range of features described in this article is available as the *semtree* package (Brandmaier, 2012b).

Note

We would like to thank Julia Delius for her helpful assistance in language and style editing.

References

Baltes, P. B., & Mayer, K. U. (Eds.), (1999). *The Berlin Aging Study: Aging from 70 to 100.* New York: Cambridge University Press.

Boker, S., Neale, M., Maes, H., Wilde, M., Spiegel, M., Brick, T., et al. (2011). OpenMx: An open source extended structural equation modeling framework. *Psychometrika, 76*(2), 306–317.

Boker, S., Neale, M., & Rausch, J. (2004). Latent differential equation modeling with multivariate multi-occasion indicators. In K. van Montfort, J. Oud, & A. Satorra (Eds.), *Recent developments on structural equation models: Theory and applications* (pp. 151–174). Dordrecht, The Netherlands: Kluwer Academic Publishers.

Brandmaier, A. M. (2012a). *Permutation distribution clustering and structural equation model trees.* Dissertation, Saarland University, Saarbrücken.

Brandmaier, A. M. (2012b). semtree: Recursive partitioning of Structural Equation Models in R [Computer software manual]. Available from http://www.brandmaier.de/semtree

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

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32.

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and regression trees.* Belmont, CA: Wadsworth International.

Brodley, C., & Utgoff, P. (1995). Multivariate decision trees. *Machine Learning, 19,* 45–77.

Chan, K., & Loh, W. (2004). Lotus: An algorithm for building accurate and comprehensible logistic regression trees. *Journal of Computational and Graphical Statistics, 13*(4), 826–852.

Dobra, A., & Gehrke, J. (2001). Bias correction in classification tree construction. In *Proceedings of the Eighteenth International Conference on Machine Learning* (pp. 90–97). San Francisco, CA: Morgan Kaufmann.

Finkbeiner, C. (1979). Estimation for the multiple factor model when data are missing. *Psychometrika, 44*(4), 409–420.

Fletcher, R. (1994). An overview of unconstrained optimization. In E. Spedicato (Ed.), *Algorithms for continuous optimization: The state of the art* (pp. 109–143). Dordrecht, The Netherlands: Kluwer Academic Publishers.

Friedman, J. (1991). Multivariate adaptive regression splines. *Annals of Statistics, 19*(1), 1–67.

Ghisletta, P., & Lindenberger, U. (2004). Static and dynamic longitudinal structural analyses of cognitive changes in old age. *Gerontology, 50*, 12–16.

Hastie, T., Tibshirani, R., & Friedman, J. (2001). *The elements of statistical learning.* Berlin: Springer.

Hawkins, D. (1999). *Firm: Formal inference-based recursive modeling* [Tech. Rep.]. Department of Applied Statistics, University of Minnesota.

Horn, J., & McArdle, J. (1992). A practical and theoretical guide to measurement invariance in aging research. *Experimental Aging Research, 18*(3), 117–144.

Hothorn, T., Hornik, K., Strobl, C., Zeileis, A., & Hothorn, M. (2011). *party: A laboratory for recursive partytioning.* Available from http://cran.r-project.org/web/packages/party/index.html

Hyafil, L., & Rivest, R. (1976). Constructing optimal binary decision trees is NP-complete. *Information Processing Letters, 5*(1), 15–17.

Ihaka, R., & Gentleman, R. (1996). R: A language for data analysis and graphics. *Journal of Computational and Graphical Statistics, 5*(3), 299–314.

Jensen, D., & Cohen, P. (2000). Multiple comparisons in induction algorithms. *Machine Learning, 38*(3), 309–338.

Jöreskog, K. G. (1969). A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika, 34*(2), 183–202.

Jöreskog, K. G. (1979). Statistical models and methods for analysis of longitudinal data. In K. Jöreskog, D. Sörbom, & M. J. (Eds.), *Advances in factor analysis and structural equation models* (pp. 129–169). Cambridge, MA: Abt Books.

Kim, H., & Loh, W. (2001). Classification trees with unbiased multiway splits. *Journal of the American Statistical Association, 96*(454), 589–604.

Lee, S., & Song, X. (2003). Maximum likelihood estimation and model comparison for mixtures of structural equation models with ignorable missing data. *Journal of Classification, 20*(2), 221–255.

Lindenberger, U., Smith, J., Mayer, K., & Baltes, P. (Eds.), (2010). *Die Berliner Altersstudie* [The Berlin Aging Study]. Berlin: Akademie Verlag.

Loh, W., & Shih, Y. (1997). Split selection methods for classification trees. *Statistica Sinica, 7*, 815–840.

Lövdén, M., Ghisletta, P., & Lindenberger, U. (2004). Cognition in the Berlin Aging Study: The first ten years. *Aging, Neuropsychology, and Cognition, 11*, 104–133.

McArdle, J. (1988). Dynamic but structural equation modeling of repeated measures data. In J. Nesselroade & R. Cattell (Eds.), *Handbook of multivariate experimental psychology* (Vol. 2, pp. 561–614). New York: Plenum Press.

McArdle, J. (2010). Some ethical issues in factor analysis. In A. Panter & S. Sterber (Eds.), *Quantitative methodology viewed through an ethical lens* (pp. 313–339). Washington, DC: American Psychological Association Press.

McArdle, J., & Aber, M. (1990). Patterns of change within latent variable structural equation modeling. In A. von Eye (Ed.), *New statistical methods in developmental research* (pp. 151–224). New York: Academic Press.

McArdle, J., & Epstein, D. (1987). Latent growth curves within developmental structural equation models. *Child Development, 58*(1), 110–133.

McArdle, J., & Hamagami, F. (2001). Latent difference score structural models for linear dynamic analyses with incomplete longitudinal data: New methods for the analysis of change. In L. M. Collins & A. G. Sayer (Eds.), *New methods for the analysis of change.* Washington, DC: American Psychological Association.

McLachlan, G., & Peel, D. (2000). *Finite mixture models.* New York: John Wiley and Sons.

Merkle, E. C., & Zeileis, A. (2011). *Generalized measurement invariance tests with application to factor analysis* (Working Paper No. 2011-09). Universität Innsbruck: Working Papers in Economics and Statistics, Research Platform Empirical and Experimental Economics. Available from http://EconPapers.RePEc.org/RePEc:inn:wpaper:2011-09

Murthy, S., Kasif, S., & Salzberg, S. (1994). A system for induction of oblique decision trees. *Journal of Artificial Intelligence Research, 2*, 1–32.

Osborne, R., & Suddick, D. (1972). A longitudinal investigation of the intellectual differentiation hypothesis. *Journal of Genetic Psychology, 121*(pt 1), 83–89.

Quinlan, J. (1986). Induction of decision trees. *Machine Learning, 1*(1), 81–106.

Quinlan, J. (1993). *C4.5: Programs for machine learning.* San Francisco, CA: Morgan Kaufmann.

Revelle, W. (2011). *psych: Procedures for psychological, psychometric, and personality research* [Computer software manual]. Evanston, Illinois. Available from http://personality-project.org/r/psych.manual.pdf (R package version 1.01.9)

Revelle, W., Wilt, J., & Rosenthal, A. (2009). Personality and cognition: The personality-cognition link. In A. Gruszka, G. Matthews, & B. Szymura (Eds.), *Handbook of individual differences in cognition: Attention, memory and executive control.* New York: Springer.

Rubin, D. (1976). Inference and missing data. *Biometrika, 63*(3), 581–592.

Sanchez, G. (2009). *PATHMOX approach: Segmentation trees in partial least squares path modeling.* Dissertation, Departament Estadistica i Investigacio Operativa. Universitat Politecnica de Catalunya.

Sanchez, G., & Aluja, T. (2012). *pathmox: Segmentation trees in partial least squares path modeling* [Computer software manual]. Available from http://CRAN.R-project.org/package=pathmox (R package version 0.1-1)

Segal, M. (1992). Tree-structured methods for longitudinal data. *Journal of the American Statistical Association, 87*(418), 407–418.

Shih, Y. (2004). A note on split selection bias in classification trees. *Computational Statistics and Data Analysis, 45*, 457–466.

Sonquist, J., & Morgan, J. (1964). *The detection of interaction effects. A report on a computer program for the selection of optimal combinations of explanatory variables* (No. 35). Ann Arbor, MI: Survey Research Centre, The Institute for Social Research, University of Michigan.

Strobl, C., Boulesteix, A., Zeileis, A., & Hothorn, T. (2007). Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics*, *8*(1), 25.

Su, X., Wang, M., & Fan, J. (2004). Maximum likelihood regression trees. *Journal of Computational and Graphical Statistics*, *13*(3), 586–598.

von Oertzen, T., Ghisletta, P., & Lindenberger, U. (2009). Simulating statistical power in latent growth curve modeling: A strategy for evaluating age-based changes in cognitive resources. In M. Crocker & J. Siekmann (Eds.), *Resource-adaptive cognitive processes* (pp. 95–117). Heidelberg: Springer.

Wechsler, D. (1949). *Wechsler Intelligence Scale for Children: Manual.* New York: Psychological Corporation.

Zeileis, A., Hothorn, T., & Hornik, K. (2008). Model-based recursive partitioning. *Journal of Computational and Graphical Statistics*, *17*(2), 492–514.

Zeileis, A., Leisch, F., Hornik, K., & Kleiber, C. (2002). strucchange: An R package for testing for structural change in linear regression models. *Journal of Statistical Software*, *7*(2), 1–38.

Zhang, H., & Singer, B. (1999). *Recursive partitioning in the health sciences.* New York: Springer Verlag.