Psychological Methods
2016, Vol. 21, No. 4, 566–582

1082-989X/16/$12.00 http://dx.doi.org/10.1037/met0000090

Theory-Guided Exploration With Structural Equation Model Forests

Andreas M. Brandmaier
Max Planck Institute for Human Development, Berlin, Germany
and Max Planck UCL Centre for Computational Psychiatry and Ageing Research, Berlin, Germany

John J. Prindle
Max Planck Institute for Human Development, Berlin, Germany

John J. McArdle
University of Southern California

Ulman Lindenberger
Max Planck Institute for Human Development, Berlin, Germany
and European University Institute, San Domenico di Fiesole, Italy

Structural equation model (SEM) trees, a combination of SEMs and decision trees, have been proposed as a data-analytic tool for theory-guided exploration of empirical data. With respect to a hypothesized model of multivariate outcomes, such trees recursively find subgroups with similar patterns of observed data. SEM trees allow for the automatic selection of variables that predict differences across individuals in specific theoretical models, for instance, differences in latent factor profiles or developmental trajectories. However, SEM trees are unstable when small variations in the data can result in different trees. As a remedy, *SEM forests*, which are ensembles of SEM trees based on resamplings of the original dataset, provide increased stability. Because large forests are less suitable for visual inspection and interpretation, aggregate measures provide researchers with hints on how to improve their models: (a) *variable importance* is based on random permutations of the out-of-bag (OOB) samples of the individual trees and quantifies, for each variable, the average reduction of uncertainty about the model-predicted distribution; and (b) *case proximity* enables researchers to perform clustering and outlier detection. We provide an overview of SEM forests and illustrate their utility in the context of cross-sectional factor models of intelligence and episodic memory. We discuss benefits and limitations, and provide advice on how and when to use SEM trees and forests in future research.

Keywords: SEM forest, model-based tree, recursive partitioning, variable importance, case proximity

The privileged unit of analysis in psychology is the individual (Nesselroade, Gerstorf, Hardy, & Ram, 2007). Nevertheless, many data-analytic approaches coarsely aggregate data and tacitly assume group-average models to hold and to be interpreted in lieu of more fine-grained and, ultimately, person-specific models. For example, when a group of persons show an average increase of performance in a learning task, this does not mean that all persons follow a pattern of change similar to this average. In fact, none of the persons may be well represented by the average trend. In a similar vein, Tucker (1966) argued that the consideration of differences instead of averages will allow us to gain more information about the nature of basic functions underlying behavior. Ever since, researchers have been questioning coarse aggregation of data across persons (e.g., Lamiell, 1981; Nesselroade & Molenaar, 1999) as the estimates of averaged effects may not be representative of any single individual. In fact, strong inference about intraindividual variation from interindividual variation is only possible under the ergodic assumption (Molenaar, 2004), which assumes that the group model represents each individual's dynamics (homogeneity) and that those dynamics have constant characteristics in time (stationarity). In the same vein, Simpson (1951) pointed out that a statistical relationship observed in a population could be reversed within subgroups that form the population. For instance, "It may be universally true that drinking coffee increases one's level of neuroticism; then it may still be the case that people who drink more coffee are less neurotic" (Borsboom, Kievit, Cervone, & Hood, 2009, p. 72). Simpson's paradox may arise whenever inferences are drawn across different explanatory levels, for example, from populations to the individual, or from cross-sectional data to intraindividual change over time (see Kievit, Frankenhuis, Waldorp, & Borsboom, 2013, for further illustrations). Hence, there still is a need for focusing on individuals or subgroups of

Andreas M. Brandmaier, Center for Lifespan Psychology, Max Planck Institute for Human Development, Berlin, Germany and Max Planck UCL Centre for Computational Psychiatry and Ageing Research, Berlin, Germany; John J. Prindle, Center for Lifespan Psychology, Max Planck Institute for Human Development; John J. McArdle, Department of Psychology, University of Southern California; Ulman Lindenberger, Center for Lifespan Psychology, Max Planck Institute for Human Development, and European University Institute, San Domenico di Fiesole, Italy.

John J. Prindle is now at the School of Social Work, University of Southern California.

We thank Sandra Düzel for assisting with the preparation and analysis of the BASE-II dataset. We would also like to thank Michael Krause and Sebastian Schröder for supporting us in growing forests on several hundred CPUs in parallel.

Correspondence concerning this article should be addressed to Andreas M. Brandmaier, Center for Lifespan Psychology, Max Planck Institute for Human Development, Lentzeallee 94, 14195 Berlin, Germany. E-mail: brandmaier@mpib-berlin.mpg.de

individuals to more accurately model individual process idiosyncrasies and similarities across persons. Particularly, in light of large-scale empirical data sets, aggregation is more likely to lead to models with low informative value about individual underlying processes as it is often difficult to expand prior hypotheses to account for the large number of potential explanatory variables.

Researchers with an awareness of the aforementioned problem face two challenges: they need to (a) determine whether there is a substantially relevant amount of heterogeneity that needs to be accounted for, and (b) account for this variability in their hypothesis-driven model. Researchers can turn to a variety of approaches to account for heterogeneity in their sample. A principled approach is to explicitly account for subgroups in the data. One can consider two distinct venues to approach the issue of heterogeneity in such multiple group models: Either heterogeneity (i.e., group membership) is assumed to be latent or assumed to be observed. If group membership is latent, researchers can rely on latent class cluster analysis (spanning latent profile analysis, latent class analysis, finite mixture models; see Vermunt & Magidson, 2002) which has the goal of finding similar data patterns when group membership is unobserved and probabilistic. That is, observed data are assumed to be drawn from a mixture of underlying probability distributions. On the other hand, if group membership is observed through a measured variable, then it can be included in the resulting model. Probably the most frequent type of analysis asks whether observed group membership predicts mean differences in a continuous outcome and comes in many flavors, such as the *t* test, paired *t* test, or the analysis of variance in its many variants. A multivariate extension encompassing these techniques as special cases are structural equation models (SEM). Multiple group SEM allow the specification of a latent variable model with differences in parameters (but also constraints if dictated by theoretical considerations) across groups. This allows for testing factor structures, growth curves, and other types of SEMs across groups, and ultimately provides a framework for directly testing the homogeneity assumption for the observed subgroups. But these approaches are faced with new challenges when confronted with large-scale empirical data. We believe that researchers typically venture out with a good theory to start with. Such a theory formalized as a statistical model typically applies to only a subset of observed variables. However, it may be unclear how much heterogeneity in a large sample could be further explained. In large-scale data sets, multiple questionnaires with even more items, genetic data, and biomarkers may form large sets of covariates with the potential to explain heterogeneity. How can a researcher decide between those variables? Which of these covariates need to be accounted for to reduce heterogeneity and improve predictive performance of a model? If a mixture model yields a statistically plausible result of dozens of subgroups, then what do these subgroups mean for the empirical questions of the researcher? How are they discernible? Which variables distinguish among them? SEM trees, a multivariate instance of decision trees, provide a formal venue to approach these questions.

Decision trees (also known as recursive partitioning methods) became prominent through the seminal work of Sonquist and Morgan (1964); Breiman, Friedman, Olshen, and Stone (1984) and Quinlan (1986), who extended the statistical paradigm to include general classification and regression problems (for an extensive overview, see McArdle, 2013; Strobl, Malley, & Tutz, 2009). They provide a data-analytic method to find subgroups in a sample with similar response patterns when group membership is observed by a subset of potential predictor variables and their interactions. This feature gave name to one of the earlier representatives of this group of methods, *automatic interaction detection*. Decision trees can be seen as a nonparametric approach to selection among a set of potential predictors to best predict univariate or multivariate outcomes. Modern decision tree approaches come in many flavors but they all share a common algorithmic theme: They partition the sample at hand into subsets that differ in their observed data patterns across groups but are similar within groups. Only partitions of the sample are considered that are described by a set of potential predictor variables. For example, if gender and age are included as potential predictors, partitions based on these variables will be considered. Once the best partition is found, the data set is permanently split into independent partitions and the algorithm proceeds recursively in each of the resulting partitions, that is, it finds the next best partition in each subset. A tree structure obtained from a dataset with age and gender as potential predictors may thus describe partitions based on either or a combination of both variables, or none at all. The result of a decision tree analysis can be visualized in a hierarchical binary tree structure. A binary tree is made of nodes, where each node has either no successors (or children) and is called a leaf or has exactly two successors and is called an inner node. The first node of the tree, which has no parent, is called the root. In a tree, each inner node corresponds to a decision, that is, to a variable and a decision rule about this variable that determines what subsets will be associated with its children. Each leaf node is describing an outcome, typically as a prediction about the outcome. To determine which node a person belongs to, one simply follows the path from a root node down to a leaf as it is determined by comparing the person's predictors with the decision nodes encountered.

In psychological research, SEMs are general modeling techniques that encompass a variety of models for testing hypotheses in cross-sectional and longitudinal studies including both observed and latent variables. In particular, latent variable SEMs allow modeling measurement errors, thereby offering greater validity and generalizability of research designs than methods purely based on observed variables (Little, Lindenberger, & Nesselroade, 1999; McArdle & Nesselroade, 2014). One special appeal of this method is the correspondence of the underlying linear equations to a graphical representation (see von Oertzen, Brandmaier, & Tsang, 2015). Conventionally, SEM is used as a confirmatory technique. Brandmaier, von Oertzen, McArdle, and Lindenberger (2013b) introduced SEM trees as a technique that combines the benefits of both SEMs and decision trees into a single method of data analysis. The method allows for theory-guided exploration of models, with structural equation modeling providing a formal framework for implementing a prior theory and the decision tree approach enabling an exploratory analysis and refinement of the initial theory. The goal of decision tree methods is essentially to discover how potential predictor variables are linked to an outcome. Trees start off by undertaking an exhaustive search among the set of potential predictors to find a predictor that best describes a partition of the sample into groups with similar patterns of the outcome. Once a partition is found, this search is recursively continued in the resulting partitions. Different evaluation functions for assessing the necessity of a further split have been proposed. For classifica-

tion of univariate outcomes, typically, GINI impurity (Breiman et al., 1984) or information gain (Quinlan, 1986, 1993) were proposed, but others are possible. SEM trees extend conventional decision trees to model-based trees that feature a parametrical model of the outcomes in each leaf instead of a prediction about a univariate outcome. Regression trees, in which each leaf model contains a regression model, are a special case of SEM trees because regression is a special case of SEM. SEM trees allow a wider variety of linear models with both observed and latent variables in leafs of the tree. Whereas conventional decision trees provide a nonparametric model of the outcomes, SEM trees fuse the nonparametric nature of decision trees with parametric SEMs as outcomes. The logic behind SEM trees is simple. Given a hypothesized SEM, the sample is recursively split to find sub groups that maximally differ with respect to the parameters of the original SEM. Typically, the hypothesized SEM is the same for each leaf, and each leaf represents a sample best described with a set of parameter estimates for the hypothesized model (but see Brandmaier, von Oertzen, McArdle, & Lindenberger, 2013a, for hybrid SEM trees that even allow different models across leafs). Similar to a mixture model analysis, SEM trees assume that the observed data is not homogeneous (that is, drawn from a single underlying probability distribution) but heterogeneous (drawn from multiple underlying probability distributions). When analyzing data with SEM trees, we assume that an initially hypothesized model is the correct model for the population but that the population is made of subpopulations, which differ in the parameters of the generating model. Thus, the population can be said to be heterogeneous with respect to the original SEM. In contrast to latent mixture models and related approaches, SEM trees not only retrieve a clustering structure of cases, but also predictors of the structure.

SEM trees use the likelihood ratio as criterion to evaluate the necessity of a split (Brandmaier, von Oertzen, McArdle, & Lindenberger, 2013b). Whenever we evaluate a particular split of the sample into multiple groups, the resulting multiple-group model (postsplit model) is nested in the original (presplit) model. If we introduced equality constraints for corresponding parameters across groups in the postsplit model, we would obtain the presplit model. This property makes the presplit model and the postsplit model nested and the likelihood ratio test applicable. Because the search across multiple potential predictors gives rise to a multiple testing problem, one cannot simply rely on the maximally selected p value but needs to adjust for multiple testing (e.g., by a simple Bonferroni correction). It can be obtained that the likelihood ratio is equivalent to scaled information gain (Brandmaier et al., 2013a) and has, as such, both statistical and information-theoretic appeal. By default, the postsplit model is obtained by simply freeing all parameters of the original SEM across groups, that is, any differences in freely estimated regressions, means, variances, and covariances of the hypothesized SEM contribute to the decision whether a partition of the sample is to be made or not. As mentioned previously, it is possible to restrict the likelihood ratio criterion to only test selected differences with respect to selected parameters that are of particular interest to the researcher. The purpose of such designs is twofold, first, allowing for more focused hypotheses, and second, strengthening the likelihood ratio test at each node by restricting the degrees of freedom for each partition tested.

A SEM tree describes an exhaustive partition of the original sample into multiple groups (each described by one of the leafs), and is ultimately a multiple group model. The important difference between multiple group models and model trees is the fact that in SEM trees group membership is not prespecified but chosen in a data-driven manner to recursively optimize the log-likelihood criterion.

The SEM tree approach allows for the detection of subgroup heterogeneity using measured variables. In principal, the method can uncover variables and their interactions that predict differences in multivariate observed data patterns according to a specified model. For example, if a theory-driven model presumes linear change in a cognitive score, trees may find a hierarchy of subgroups with different change trends due to training dosage.

Conceptualizing different change profiles is a crucial ontologic and diagnostic activity in psychological research. For example, Muthén (2004) explored subgroups of longitudinal trajectories in high school mathematics achievement, or Josefsson, de Luna, Pudas, Nilsson, and Nyberg (2012) identified differential trajectories of episodic memory development. SEM trees and forests allow for the detection of covariate-specific subgroups, for example, in regression models, factor analytic models (Jöreskog, 1969), autoregressive models (Jöreskog, 1970), latent growth curve models (McArdle & Epstein, 1987), latent change score models (McArdle, 2009; McArdle & Hamagami, 2001), or latent differential equations (Boker, Neale, & Rausch, 2004). However, a concern of trees is their potential instability (e.g., Berk, 2006; Hastie, Tibshirani, & Friedman, 2001). In each inner node of a tree, both the associated predictor variable and its split point are chosen to be locally optimal and, thus, can easily be influenced by small perturbations of the sample at hand. A slightly different choice of a split point may lead to a different choice of the subsequent split in the children of a node; in this way, small perturbations are typically magnified down the tree. Instability is usually observed when predictors are highly correlated or variables are equally informative about different subpopulations.

In machine learning, *ensemble methods* were proposed to improve the robustness and accuracy of individual models. An *ensemble* refers to a set of predictive models that are, typically, each based on a random sample of the original data set. Ensemble methods are metalearning algorithms specifying (a) the sampling scheme for generating the data sets for the individual models, and (b) a combination scheme for aggregating the predictions of the individual models into a final prediction. Currently, the most widely known ensemble method for decision trees is random forest. The rationale of random forests is to rely on a variety of trees each based on a random sample of the original data to account for the potential instability of each individual tree. First, *bootstrap aggregating* (bagging) has been proposed as a means to create forests of decision trees. Bagging generates random samples of the original data by sampling cases uniformly and with replacement (bootstrapping). Bagged forests make predictions by aggregating predictions of the individual trees (Breiman, 1996), that is, by predicting a continuous outcome as the average over the individual trees' predictions, or by predicting a dichotomous outcome with the majority vote over the individual tree's predictions. As an extension to bagging, Breiman (2001a) suggested to randomly draw both cases and predictors for each tree in a random forest. The former disadvantage of trees, their susceptibility to random

fluctuations in the sample, becomes an advantage in an ensemble of trees. Increasing diversity can allow an ensemble to represent a better approximation to the potentially complex true partition of the sample. If cases and predictors are randomly sampled, then the diversity of the resulting trees is increased. Formal results show that increased diversity is beneficial for the performance of ensemble methods (Bühlmann & Yu, 2002). In an empirical comparison of various classification methods over 176 data sets, random forests outperformed other classification approaches including generalized linear models and support vector machines (Fernández-Delgado, Cernadas, Barro, & Amorim, 2014).

In this article, we extend SEM trees to SEM forests following the seminal work on random forests by Breiman (2001a). We present the procedure for growing a forest of SEM trees and describe two aggregate measures that allow researchers to obtain useful information about heterogeneity in their datasets: (a) *variable importance*, which quantifies the extent to which variables predict differences with respect to the initial SEM, and (b) *case proximity*, which enables researchers to perform case-based clustering based on a measure of similarity in predictor space. The benefit of forests is to provide more robust and effective measures of yet unmodeled[1] information that may help to explain observed heterogeneity. Below, we will present two applications to illustrate the approach, (a) a factor model of intelligence, and (b) a factor model of episodic memory. The first data set may not be seen as particular "big" but was chosen to complement our earlier tree analyses (Brandmaier et al., 2013b) and to highlight the added benefits of forest analyses. The second analysis is included to provide a data set that may be considered closer to a big data type of application.

The SEM forest program is freely available to researchers. We have implemented SEM forests within the semtree package (Brandmaier, 2015; see also Brandmaier et al., 2013b) for the statistical programming language R (R Core Team, 2013) including computation and plotting facilities for variable importance and case-based proximity. Researchers with access to computing networks have the possibility to profit from parallel growing of trees within a forest.

Method

SEM forests extend single SEM trees to ensembles of SEM trees, just as random forests extend decision trees to ensembles of decision trees. SEM trees are hierarchical structures of decision rules that describe differences between recursive partitions of a sample with respect to a SEM. More explicitly, SEM trees extend univariate decision trees to decision trees modeling multivariate outcomes.

A single SEM tree, potentially one of many[2] in a SEM forest, is created as follows (see Brandmaier et al., 2013b for a detailed description of nonrandom SEM trees):

Let M be a hypothesized SEM and let D be a data set containing variables in M and further variables as potential predictors of differences of observations in D with respect to the model M.

1. Among all potential predictors in D, randomly sample a subset of c candidate predictors.

2. Choose variable v among all candidate predictors that is most informative about heterogeneity with respect to M, that is, choose the variable that finds the largest difference between the resulting partitions of the sample. This is formalized as choosing the variable maximizing a likelihood ratio criterion.

3. Stop searching for further splits if a stopping criterion is reached. Stopping criteria may include (a) a minimum number of cases is reached in the current partition such that fitting a SEM is unreasonable; (b) a user-specified maximum height of the tree is reached, or (c) a statistical criterion is reached, for instance, none of the potential predictors has a significant p value when assessing the splits.

4. Else, permanently partition the dataset according to the split point of v with the largest likelihood ratio, create two new nodes of the tree, and restart this algorithm with each of the resulting partitions as D.

The SEM forest algorithm is as follows:

Let t be the number of trees in a forest, also called the forest size or ensemble size. For $i = 1, \ldots, t$, create data set D_i^{train} by randomly sampling cases (sampling is done by bootstrapping or preferably subsampling; Strobl, Boulesteix, Zeileis, & Hothorn, 2007). Assemble the remaining cases for each i in the *out-of-bag sample* D_i^{OOB}. Then grow a tree for each resampled dataset D_i^{train}.

Only a subset of all potential predictors, which we refer to as candidate predictors, is tested at each node in a random forest. The size of the set of candidate predictors, c, at each node is typically heuristically chosen as either 1, 2, $c = \lfloor \log_2(c) + 1 \rfloor$, $c = \sqrt{m}$, or $c = m/3$ (e.g., Breiman & Cutler, 2014; Verikas, Gelzinis, & Bacauskiene, 2011) with m being the total number of potential predictors. Note that if c equals m, then the random forest algorithm reduces to bootstrap aggregating (also referred to as bagging) of trees since only cases but not variables are subsampled. The predictive accuracy[3] of forests is generally higher with lower correlations among trees (Breiman, 2001a). A smaller c increases the variability between trees, and reduces the chance of suppression, that is, the chance that predictor influences remain undetected through the effect of variables with a stronger influence. On the other hand, a very small c reduces the chance to generate enough trees containing true important predictors and to detect higher-order interactions (Díaz-Uriarte & De Andres, 2006). The parameter c is a so-called hyperparameter[4] that can either be set manually or be automatically tuned, for example, by employing a hold-out data set or cross-validation; however, it seems a less critical hyperparameter than in many other analysis approaches (Strobl et al., 2009). An empirical test of c's criticality is the parallel construction of multiple SEM forests with slight variations in c and possibly the forest size t. If results of aggregate statistics as described below change considerably, hyperparameters need to be adapted.

Random sampling is typically either performed by bootstrapping or by subsampling (see Strobl et al., 2007). In a bootstrap

[1] Unmodeled in the sense that this information is not part of the initial theory-driven SEM.

[2] Details about the number of trees will be discussed later.

[3] Explained variance is a measure of relative gains in predictive accuracy.

[4] Hyperparameters are parameters that are not adjusted by the model fitting procedure but instead either set manually or adjusted by an external criterion.

sample, a new sample is obtained by drawing cases with replacement with the same sample size of the original data set. A subsampled data set is obtained by drawing a smaller data set than the original data set without replacement. In both instances, undrawn cases remain and form the out-of-bag sample that proves useful as a validation sample for hypotheses that were created with the former random sample. Because randomness is an essential part of the method, it is necessary to manually control and store the seed of the computer's random number generator to make analyses reproducible.

Partitioning of ordinal, continuous, and categorical variables is performed by internally dichotomizing each variable type (Brandmaier et al., 2013b). This approach is known as an *exhaustive split search* (Quinlan, 1993). Exhaustive split search simplifies the resulting tree as it enforces binary trees as outcome, that is, trees that branch at each node into either none or two children. Furthermore, this approach allows for testing effects of potential predictors on the SEM independent of any monotonic transformation, such as, normalization, logarithmic transformation, or polynomial transformation. Still, the dichotomization of categorical variables is exponential in the number of categories and prohibitive already for small numbers of categories. For example, a variable with 12 categories requires performing more than 4,000 model estimations in each node of the tree. To alleviate this problem, only a random sample of dichotomized candidate splits is examined (Breiman, 2001a). We adopted this procedure such that, when $\log_2(n) < k$ for a variable with k categories and n remaining cases, only $\log(n)$ split candidates from a random uniform distribution are chosen. Asymptotically, the search for the optimal split in categorical variables is then in the order of magnitude of the search in continuous variables.

Variable Importance

Measures of variable importance quantify the impact a variable has on the overall prediction of a response, which in SEM forests typically takes the form of a multivariate means and covariance structure. For instance, counting the frequency with which a variable was selected in a forest is a naive implementation of variable importance. However, this does not necessarily reflect the impact a variable has in the prediction of a model-based response. Instead, recent variable importance measures are based on permutation accuracy importance. The intuition being if an important predictor is randomly permuted, and, thus, its functional relation with the model-predicted distribution is broken, then a considerable decrease of the model's goodness-of-fit for the scrambled data is expected. In other words, permuting the variable is seen as a proxy for removing the effect of that variable. By averaging the decrease in fitness across all trees of the forest on the out-of-bag (OOB) samples, an estimate of importance is obtained for each variable. For classification trees, the Gini criterion is often employed as an optimization criterion (Loh & Shih, 1997). In SEM forests, Gini importance is replaced with log-likelihood importance, and, thus, the decrease of log-likelihood is averaged over the forest to obtain variable importance estimates. As described above, randomly sampling from both cases and predictors in the process of growing trees allows potentially influential variables to play out their effects in different, randomly sampled interactions without constantly being outrivaled by stronger competitors. The computation of variable importance in SEM forests follows from variable importance computation in conventional random forests (Breiman, 2001a):

For each potential predictor c,

1. Calculate the log-likelihood of the OOB sample for each tree, $LL(\boldsymbol{D}_i^{OOB} \mid T_i)$.
2. Obtain a scrambled OOB sample $\tilde{D}_i^{OOB}$ by randomly permuting the column of $\boldsymbol{D}_i^{OOB}$ corresponding to c.
3. Calculate the likelihood of the scrambled OOB samples, $LL(\tilde{D}_i^{OOB} \mid T_i)$.
4. Obtain the estimate of variable importance by averaging over the log-likelihood-ratios, $\psi_c = \frac{1}{t}\sum_{i=1}^{t} LL(D_i^{OOB} \mid T_i) - LL(\tilde{D}_i^{OOB} \mid T_i)$.

The above algorithm yields variable importance as an average absolute increase in model misfit due to randomization of each variable. We recommend reporting importance as increase in model misfit relative to baseline fit because it is independent of the sample size. Alternatively, Breiman and Cutler (2014) calculate standardized importance scores (z-scores) from the raw importance scores ψ_i as $\Psi_i = \frac{\psi_i}{\sigma_i}\sqrt{t}$ with σ_i being the standard deviation across each of the ψ_i. This approach allows for a significance test of variable importance but is overpowered, if not meaningless, due to the typically large number of trees in a forest (Strobl & Zeileis, 2008). Hapfelmeier and Ulm (2013) reviewed different approaches to variable selection based on variable importance. They calculate p values for each variable derived from an empirical null distribution in a permutation test framework. When mixed variable types are contained in a dataset, the use of an unbiased variable selection criterion is recommended when SEM forests are estimated. Otherwise importance analyses were biased, typically in favor of variables with larger number of categories; under the null hypothesis, continuous variables were typically selected over ordinal variables, and among ordinal variables, those with more categories were selected over those with fewer. SEM trees offer unbiased variable selection methods (Brandmaier et al., 2013b) and, thus, remove this bias.

Proximity Measures

Proximity measures quantify the pairwise similarities of cases in a sample and, thus, provide a measure of the internal structure of the data. The calculation of proximity follows the ideas of Breiman and Cutler (2014). Clustering is a technique to find hidden structure in data. Clustering based on a proximity matrix can be used to uncover heterogeneity in the dataset, detect outliers, find prototypes, or impute missing values. The intuition behind case-based proximity is that similar cases should end up in the same leafs of trees across the forest. Vice versa, each pair of cases in a terminal node shares similar values of the variables along the path back to the root node. Because the variables were chosen to locally maximize the information about the model predictions, the similarity measures take into account the differential weighting of importance for the prediction as it is encoded in the forest. The proximity matrix is symmetric, and has rows and columns each correspond-

ing to one case of the original data. To compute the proximity matrix, we count the relative frequency of each pair of cases appearing in the same terminal nodes across the forest and store this result in the corresponding entry of the proximity matrix. This notion of proximity in covariate space is then based on the relevance of variables to explain differences in the model-predicted distribution. Particularly, variables that are less important contribute less to the proximity. By construction, the proximity matrix is symmetric and bounded by 1. For large data sets, the proximity matrix is too large to be useful for inspection because the contained information grows quadratically with the number of cases. Beyond clustering, projections of the proximities on coordinates in a lower number of dimensions, for example, via multidimensional scaling (MDS), can shed light on the case-by-case similarity in a dataset. Given a dissimilarity matrix, MDS arranges objects in a low-dimensional space such that the projected proximities have maximal fidelity to the original proximities. This is usually achieved by a principal components analysis of the dissimilarity matrix. The resulting components are then referred to as principal coordinates and the coordinate system is defined by principal axes. Then, the cases can be visually represented by plotting their principal coordinates. Because MDS requires a dissimilarity matrix, the proximity matrix, P, needs to be transformed into a dissimilarity matrix, D, before MDS can be applied:

$$D = 1 - P$$

A heuristic measure of "outlyingness" or "novelty" for classification was proposed by Breiman and Cutler (2014) as the reciprocal of the sum of squared proximities between a case and all other cases conditional on the same class. In settings with multivariate continuous responses, no class information is available and the novelty is calculated as unconditional average: Let P be the proximity matrix derived as described above. The novelty of an observation i is its average dissimilarity to all other cases:

$$d_i = N \setminus \sum_j^N P^2(i, j)$$

Breiman and Cutler (2014) suggested normalizing this score using robust statistics for location (median; MED) and dispersion (median of absolute deviation; MAD):

$$\text{MED} = \text{median}_i(d_i)$$

$$\text{MAD} = \text{median}_i(|d_i - \text{MED}|)$$

$$\tilde{d}_i = \frac{d_i - \text{MED}}{\text{MAD}}$$

Intuitively, novelty of a case is large when its proximity to all other cases is on average small. In other words, the less often a case is consistently allocated to similar cases in the same leaf node across the forest, the greater its novelty value.

Proximity and importance complement each other. Variable importance is a nonparametric estimator of the information a variable can convey about the model-predicted distribution in the interaction with other potential predictors. The aggregation of importances across the trees hides the partitional information of single trees. Case-based proximity recovers parts of this information by defining a similarity between cases that is implicitly weighted by the importance of their variables in the forest. In addition to the predictive ranking of variable importance, proximity may provide structural insights into potential extreme groups and outliers.

Application

To illustrate the utility of SEM forests, we present analyses of two empirical data sets, of which one was investigated with SEM trees before (Brandmaier et al., 2013b). In Appendix A, we provide a worked R example to demonstrate how SEM Forests can be practically used.

Factor Analysis of the Wechsler Adult Intelligence–Revised

We investigated variable importance for a cross-sectional factor model of verbal cognitive ability and showed how model modification based on importance may be performed. The underlying research question is concerned with exploring how we can better explain individual differences in verbal ability in a cross-sectional sample. This sample of the Wechsler Adult Intelligence–Revised (WAIS–R) was previously analyzed by others (e.g., Horn & McArdle, 1992; McArdle & Hamagami, 1992; McArdle & Prescott, 1992). The sample was collected during 1976 and 1980 by the Psychological Corporation and includes the scores of N = 1,880 individuals (age from 16 to 74 years) on a total of 11 WAIS–R subscales. A rich set of demographic variables is available for this dataset, which we selected as potential predictors, including age group (in nine ordinal categories, hence eight implied dichotomous variables), geographical information about the place of residence (four nominal categories, hence seven implied variables), urban/rural place of residence (dichotomous), born in the U.S. (dichotomous), marital status (again four nominal categories, hence seven variables), race (three nominal categories), Hispanic heritage (dichotomous), handedness (dichotomous), education (six ordinal categories, hence five variables), occupation (six nominal categories, hence 31 variables), sex (dichotomous) and birth order (nine ordinal categories, hence eight variables).

Following Brandmaier et al. (2013b), we set up a single-factor model that hypothesizes one latent factor F for verbal cognitive ability, $\boldsymbol{Y} = \Lambda F + \boldsymbol{\epsilon}$, with $\Lambda = (1, \lambda_2, \lambda_3, \lambda_4)$, $\boldsymbol{Y}$ being the vector of indicator scores, and $\boldsymbol{\epsilon}$ the vector of residuals $(\epsilon_1, \epsilon_2, \epsilon_3, \epsilon_4)$. The four indicator variables for verbal performance were information, comprehension, similarities, and vocabulary. In addition to fixing the loading of information, the expectation of this variable was constrained to 0, allowing the mean and variance of the factor to be estimated.

We randomly split the sample (N = 1,880) in two halves of equal size, a training set and a hold-out set, to allow confirmation of a forest-based exploration. The forest analysis was performed on the training set only. The resulting variable importance served as basis of a modification of the original SEM, which was in turn evaluated on the hold-out set.

A forest analysis with 1,000 trees (see Figure 1), subsampling of cases across the tree, and $m = [\log_2(12)] = 4$ (the dataset contains 12 potential predictors, so we randomly sample four split candidates at each node) yields "education" as the most important variable, with an average absolute increase of −2LL of 363.22, which corresponds to an average decrease in loglikelihood of about 12.5%. To reiterate, this is the drop in likelihood averaged

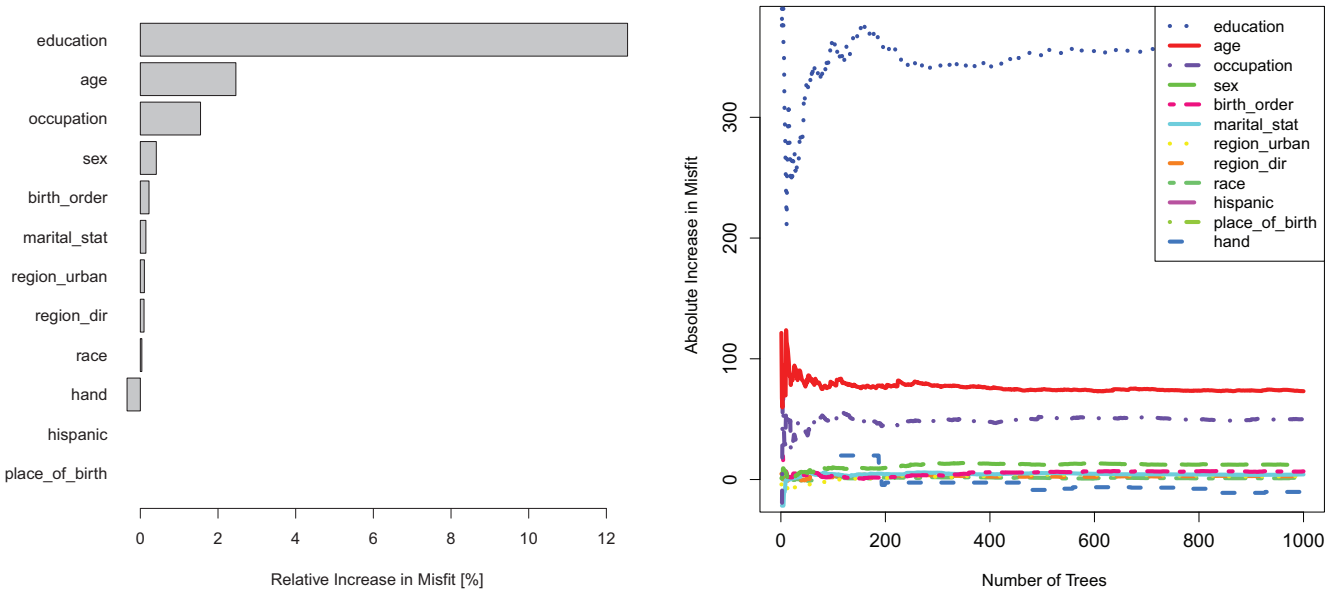

Figure 1. Left: A bar chart of variable importance for the WAIS–R factor model, quantified as average increase in model misfit due to randomization. Right: Convergence behavior of the absolute importance (*y*-axis) over the number of trees (*x*-axis). See the online article for the color version of this figure.

over the forest when the variable "education" is randomly permuted. Second and third runner-up are "age" with only a fifth of the effect of "education" and "occupation" with 13.7%, respectively. Based on this result, we propose a modification of the model to include the effect of the most important predictor, "education," on the factor structure. Because "education" was coded as an ordinal variable,[5] we created a single tree with only a single split at the root node to determine the maximally informative split into two groups. The resulting tree partitioned the sample into participants who graduated from high school (12+ years) or not (0–11 years) as the most informative split ($\chi^2 = 372.16$, $df = 12$, $p < .001$). We hypothesize education to predict differences in mean verbal performance and, thus, created a variation of the previous factor model including high school graduation as an exogenous predictor of the latent verbal ability. If the null hypothesis were true that education were uncorrelated with our model variables, this model should have reasonable fitness when the influence of graduation was restricted to zero. The model fit including the zero constraint was unacceptable (RMSEA = 0.24, CFI = 0.89, SRMR = 0.25). When freeing the regression of verbal ability on graduation, model fit was fine (RMSEA = 0.069, CFI = 0.99, SRMR = 0.01). We formally tested the inclusion of the regression with a likelihood ratio test and could significantly reject the removal of the regression path ($\chi^2 = 317.6$, $df = 1$, $p < 0.001$).

This model modification serves as an example of how a SEM forest can inform the amelioration of a theory-driven, explanatory model. Note that the strength of variable importance is the quantification of the overall predictive effect in the interaction with all other predictors whereas our proposed modification was limited to the inclusion of a single variable. For an example of a model modification using the continuous variable "age," see McArdle (1994) or interactions of "education" and "age," see McArdle and Prescott (1992). Previous studies have also shown the verbal IQ construct to be related to education, age, and sex differences (Kaufman, Reynolds, & McLean, 1989; Reynolds, Chastain, Kaufman, & McLean, 1987).

The single SEM tree reported by Brandmaier et al. (2013b) provided one way to look at potential predictors and their interaction but was already at the verge of interpretability due to its large number of splits. The authors concluded that the variable "education" was the most important since the first two splits in a single-tree analysis contained this variable; these splits separated the extreme groups with very high and very low education from the midfield that was further partitioned by the tree. Further partitions in the single-tree analysis were made with respect to the variable "occupation." A more rigorous SEM forest analysis supported the observed importances in the single-tree solution but added a robust quantification of their relative strength.

As further means of exploring structure in the data set, we calculated proximity between each pair of the 940 training cases and projected each case onto the first two principal coordinates of their dissimilarity matrix (Gower, 1966), that is, a two-dimensional representation that minimizes loss of information. As a result, we obtain a two-cluster structure that reflects the most important variables from the preceding analysis (see Figure 2). The cluster structure along the vertical principal axis clearly shows two sets of three clusters discriminating education (as shown by the symbols encoding level of education). The horizontal principal axis differentiates between people in and out of the workforce. Within each cluster a third principal axis encodes age as shown by the color-coding of the symbols. We followed up on this result by applying

[5] Years of education were coded: 1 ← 0–7 years; 2 ← 8 years; 3 ← 9–11 years; 4 ← 12 years; 5 ← 13–15 years; 6 ← 16+ years.

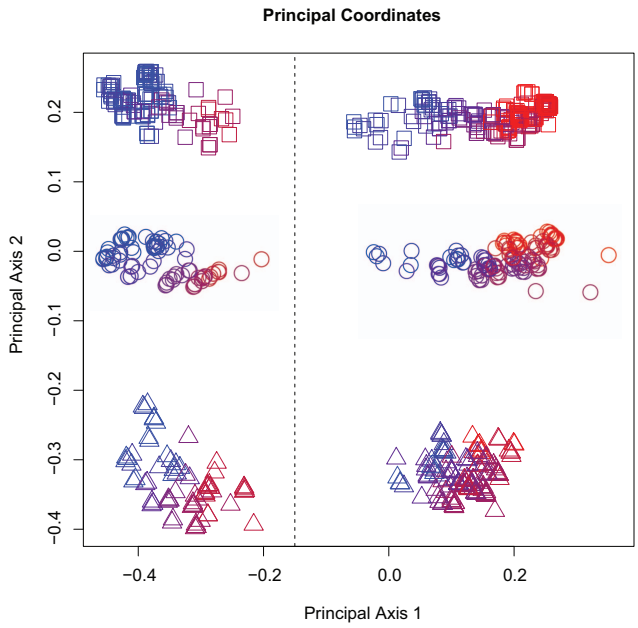

Figure 2. Proximity of participants with respect to the WAIS–R verbal factor model. The plot shows participants on the First two principal coordinates of the proximity matrix. The color gradient represents age from 16 (red) to 74 (blue). Levels of education are coded as follows: 0–8 years (square), 9–11 years (circle), 12 or more years (triangle). The dashed line separates people with respect to their employment status; cases on the left are unemployed, cases on the right are in the labor force. See the online article for the color version of this figure.

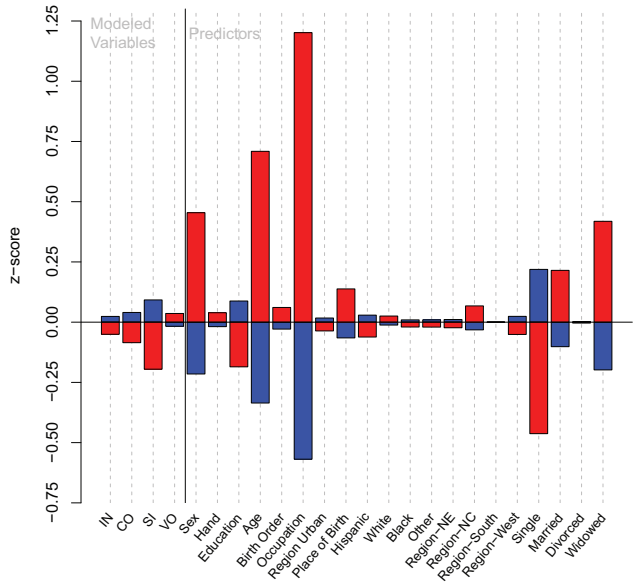

Figure 3. Group-wise predictor means on a z-scale for two groups derived by k-medoids clustering of the proximity matrix. Categorical variables marital status and geographic region (direction) are represented in dummy coding. See the online article for the color version of this figure.

a *k*-medoids[6] clustering algorithm to the proximity matrix. The number of clusters was chosen in a data-driven way by choosing *k* to maximize silhouette width (Rousseeuw, 1987), a measure of clustering consistency, which relates the average distance of each case to members of its cluster to the average distance of each case to all members of the next closest cluster. From these two clusters, we calculated the means of each predictor and of the modeled variables on a z-scale. The result is shown in Figure 3. The plot can serve as a means to inspect the homogeneity of the predictor structure across participants. As can be seen from the average deviation of the modeled variables, one homogeneous cluster that was identified here (shown in blue color) comprises mostly unemployed (99.7% of the cluster) older (m = 54 years) women (73% of the cluster) whereas the other larger cluster representing the remaining sample does not conspicuously deviate from the whole sample average. The choice to interpret two clusters was based on a data-driven, exploratory step and may be repeated with larger number of clusters for further exploration.

Exploring Predictors of Differences in a Factor of Episodic Memory

As a further practical example of how trees and forests can be used to explore heterogeneity in empirical data, we analyzed data from the Berlin Aging Study II (BASE-II; Bertram et al., 2014). Here, the exploratory analysis focused on the question of how much a set of diverse predictors spanning psychosocial, demographic, and health-related indicators may, potentially in their interaction, help to predict individual differences in episodic memory. BASE-II is an interdisciplinary study investigating physical, cognitive, and social conditions associated with successful aging. The analysis was based on the complete sample of 2,463 participants, of which about one quarter were between 20- and 35-years-old and three quarters were between 60- and 80-years-old. We analyzed cross-sectional data from the first available assessment, which had started in 2009. We created a single factor representing episodic memory. The construct was indicated by four items: (a) the verbal learning and memory test assessing auditory verbal learning. The sum of items recalled over five trials constituted the item score; (b) the face profession task assessing associative binding on the basis of recognition of incidentally encoded face-profession pairs. Corrected hit rates for rearranged face-profession pairs were used as indicator; (c) the object location memory task assessing object-location memory with 12 colored photographs arranged on a 6 × 6 grid. The sum of correct placements was used as the observed score; and (d) the scene encoding task assessing the ability of incidental scene encoding. A delayed recognition hit-rate was used as manifest variable. A more detailed description of these tasks was given by Düzel et al. (2016). The overall model had a good fit (CFI = 0.997; RMSEA = 0.032; p_{RMSEA} = 0.797). From the available covariates, we ad hoc selected a subset representing diverse indicators spanning psychosocial, demographic, and health-related indicators in an ad hoc manner. Among these, we chose to include age group (young/old), sex, years of education, marital status, and number of children; self-reports on sleep quality (day and night), smoking, healthy eating, the frequency of

[6] Medoids are similar to centroids (or geometric centers), only that they always are members of the data set whereas centroids may lie between members.

social contacts (each with partner, relatives, friends, acquaintances, and neighbors), frequency of communication via telephone (each with partner, relatives, friends, acquaintances, and neighbors), the number of close friends, assessment of the current financial situation, time pressure, life satisfaction, optimism, goal engagement, and control beliefs (internal/external/other). Health-related indicators included occurrence of diseases (diabetes, asthma, coronary diseases, cancer, stroke, migraine, hypertension, depression, dementia, joint diseases, backpain, and sleeping disorders), and reports of sports activity and physical limitations in day-to-day work; we included positive and negative affect from the PANAS (Watson, Clark, & Tellegen, 1988), the big five personality traits (Lang, John, Lüdtke, Schupp, & Wagner, 2011), and the TICS (Schulz & Schlotz, 1999) for stress. From the subjective health questionnaire of the German Socioeconomic Panel (see Böckenhoff et al., 2013), we added self-rated overall wellbeing and health satisfaction, and further items of self-rated satisfaction with sleep, work, health, household tasks, and household income. All items were rated on 11-point Likert scales. We added three questionnaire items of loneliness (missing company of others, feeling left out, and feeling isolated). Lastly, we added dichotomous indicators of life events including severe injury or disease of self, severe injury or disease of the partner, death of a family member, divorce, severe conflict, financial burden, responsibility for a person in need of care, and relocation. This led to a data set with 73 potential predictors to explain heterogeneity with respect to a latent factor of episodic working memory.

With the single-factor model as outcome, a SEM forest was grown. We set $m = \lceil \log_2(73) \rceil = 7$ (the dataset contains 73 potential predictors, so we randomly sample seven split candidates at each node), a forest size of 2,000, and computed variable importance. Results are plotted in Figure 4. Unsurprisingly, age was by far the most influential effect on the episodic memory factor. This is in line with the life span perspective on the profound and continuous changes in EM in the sense of a decrease starting in middle adulthood with accelerating decline in very old age (cf. Shing et al., 2010; Singer, Verhaeghen, Ghisletta, Lindenberger, & Baltes, 2003). The second most influential variable—with an estimated effect of a tenth of the aging effect—were work satisfaction, which may be seen as proxy for both general wellbeing and stress, and stress was found to negatively impact memory performance (VonDras, Powless, Olson, Wheeler, & Snudden, 2005; Wolf, 2009). The appearance of relocation on a similar rank is surprising at first; one may hypothesize that relocation also is a major stressor in life. Alternatively, relocation may indicate that individuals are no longer able to live independently due to increasing frailty. Following up, we find hypertension. Hypertension is associated with impairments in cognitive functions in older adults (Bender, Daugherty, & Raz, 2013; Raz, Rodrigue, & Acker, 2003). A potential pathway links hypertension as an important cause of cerebrovascular disease (CVD) associated with mild cognitive impairment, defined as episodic memory impairment beyond the degree as seen in healthy aging (Nordahl et al., 2005). Hypertension and diabetes, both among the top predictors found by the forest, are primary risk factors for CVD, which stresses the importance to further research this proposed pathway. Physical limitations in day-to-day work is likely predictive in its role as proxy for overall health of an individual. Further predictors were education, the occurrence of joint diseases, then a indicators describing the personal family situation (being in a partnership), and back pain as a further disease-related predictor (with only a fortieth of the importance of age group). For sake of brevity, we omit the exact ordering of the remaining predictors. In summary, we conclude the forests successfully retrieved predictors previously found to be associated with impaired memory functioning or verbal intelligence.

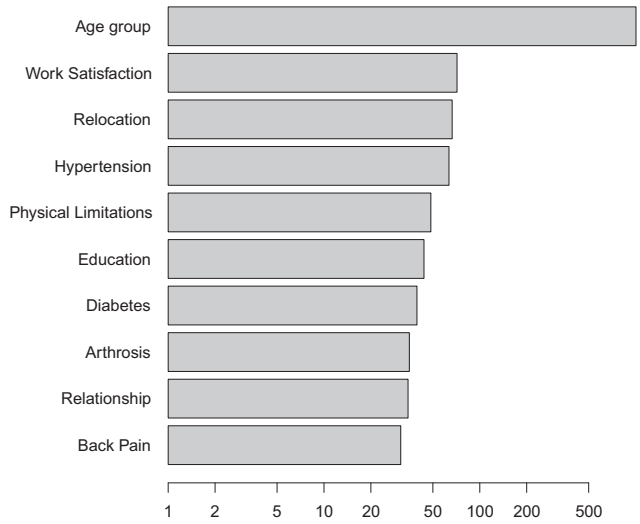

Figure 4. Top 10 variable importance estimates for an episodic memory factor in the BASE-II study plotted on a log-scale. From most important to least important: Age group (young/old), work satisfaction (11-point scale), relocation (yes/no), hypertension (yes/no), physical limitations in day-to-day work, education (in years), diabetes (yes/no), arthrosis (yes/no), being in a relationship (yes/no), back pain (yes/no).

Discussion

SEM forests constitute a hybrid of two modeling cultures (see Breiman, 2001b; Shmueli, 2010, for perspectives on the differences): (a) structural equation modeling, which is a theory-driven, explanatory modeling approach that yields interpretable models with known statistical properties; and (b) random forest, which is a data-driven, predictive modeling technique allowing the retrieval of important subsets from large numbers of predictors with potentially complex interactions. Despite the fact that predictive modeling techniques are largely ignored for scientific inquiries, both explanatory and predictive modeling are useful for generating and testing theories (Shmueli, 2010). Predictive modeling complements explanatory modeling on multiple dimensions. As demonstrated in the SEM forest approach, predictive modeling can suggest improvements to explanatory models. In this way, predictive modeling helps to generate new hypotheses, particularly in large and complex data sets in which it is difficult to hypothesize (or difficult to winnow in on which variables to focus in on). Furthermore, predictive modeling approaches can be employed as a baseline of predictive accuracy that may serve as a benchmark for purely explanatory models: If predictive modeling is more accurate than some explanatory model, the explanatory model is likely to have room for improvement.

Benefits and Limitations

A key feature of both SEMs and simple decision trees is their readability and interpretability. In trees, feature interactions and resulting partitions can be easily visualized. In SEMs, the hypothesized structure of causes and effects between variables can be represented graphically (e.g., see von Oertzen et al., 2015). In contrast, as Breiman (2001a) states, forests are "impenetrable as far as simple interpretation of its mechanisms go" (p. 23). The straightforward interpretation of a single tree is lost in the forest. Particularly, forests do not provide a straightforward hierarchical clustering of the participants into homogeneous subgroups. However, the variable importance estimate yields a structured approach to aggregate information about previously unmodeled variables across the trees. In the endeavor to explain phenomena, the question often arises which variables should be controlled for, respectively, what variables can provide predictive information; and how strong is their influence on the model of interest. SEM forests provide a starting point for researchers to address these questions in a rigorous and robust manner.

Variable importance and proximity measures provide a way to efficiently search empirical data for sources of variability. Our empirical results are supported by the interactions of demographic characteristics as found by traditional analysis techniques for the WAIS–R verbal factor, but they also go beyond the recovery of known associations. Whereas in SEM trees variables compete with one another for model impact, SEM forests assess competing variables' importances by the controlled utilization of randomness and, thus, provide an unbiased estimate of marginal importance for model fitness.

Still, SEM forests impose high computational burdens, and a forest analysis may thus be time demanding. For large sets of predictor variables, the number of model optimization runs in a SEM tree depends linearly on the number of potential split points and linearly on the number of observations (Brandmaier et al., 2013b). In trees, all predictors are tested at each level, whereas, in forests, random sampling of the predictors will drastically decrease the number of tests and, thus, decrease the number of potential split points, leading to a reduced asymptotic time complexity. Under certain circumstances, typically when a large number of predictors are chosen, a forest analysis can even be faster than a single tree analysis. Still, forests are typically created with hundreds to thousands of members, which in our experience prohibits computation on desktop machines for most models. However, models featuring a small covariance matrix of size 4×4 and an ensemble with 1,000 members may be generated in less than 1 hr on a standard desktop computer. The ideal size of a forest depends on the number of predictors, their interactions, the heterogeneity of the data, and the complexity of the model. Currently, good guidelines for forest size are still missing and, when computational resources are sparse, researchers are advised to iteratively increase forest size until forest results empirically stabilize. Note that forests need not be regrown but can be grown by simply adding more trees.

Choosing the hyperparameters (i.e., parameters of the sampling procedure) remains a crucial decision when estimating variable importance using SEM forests. Typically, this entails selecting the resampling procedure (bootstrapping or subsampling), selecting the number of variables that are sampled at each level, and the number of trees in the forest. Currently, we heuristically rely on evidence from the research on (conditional) random forests. Further simulation work is needed to outline optimal hyperparameters for creating SEM forests. Currently, we set the number of trees to 2,000 and empirically examine whether variable importance converges as the number of trees increases to decide whether a larger forest is needed. We heuristically set the number of candidate predictors to the logarithm of the total number of predictors and use subsampling as resampling scheme. As noted earlier, the choice of hyperparameters may influence the results of a forest. For example, when choosing the number of candidate variables, there is a trade-off between diversity and stability: A low number of candidate variables leads to more diverse trees at the cost of important predictors having a lower chance to enter any single tree. Other parameters that directly influence the likelihood ratio test, which ultimately underlies the variable selection mechanism, such as missingness or group imbalance in categorical variables, may bias variable selection and affect the power to detect important variables. Similarly, the measure of proximity is influenced not only by the choice of hyperparameters but also by the number of variables and their relative importance. Future work needs to address these questions from theoretical and empirical perspectives across many types of SEMs ranging from cross-sectional factor models to coupled latent differential equations.

As a tool for improving a theory-driven SEM, variable importance grants no free lunch. It is commonly accepted that theory construction and refinement benefit from measures that specify the relationship of relevant predictors to outcome variables (e.g., in the sense of regression), the relationship among relevant predictors (e.g., in the sense of moderation), or both. Variable importance falls short of either mark but offers two other benefits in turn. First, variable importance integrates the predictiveness of a given variable into a single score, and it does so in a way that includes all of its interactions with other predictors across all trees of the forest. It should be noted that alternative ways to derive such measures are available. For instance, Strobl, Boulesteix, Kneib, Augustin, and Zeileis (2008) discussed conditional instead of marginal variable importance, and Hapfelmeier, Hothorn, Ulm, and Strobl (2014) introduced an alternative variable importance scheme that replaces OOB scrambling by randomly distributing cases to each node while holding the ratio of cases over child nodes constant. In the presence of missing values of variables, this scheme was found to be superior over standard variable importance. The second benefit of variable importance is related to the first: Variable importance expresses the expected decrease in uncertainty for the entire outcome model instead of being restricted to a specific variable or relation in that model. Taken together, then, variable importance conveys useful summary information about predictor variables in relation to an outcome model. Nevertheless, it is difficult to indicate how exactly this information should be used to guide theory development. How can we best translate the results of a forest analysis, and the summary information conveyed by variable importance, into an improved parametric model, and ultimately build a better scientific theory? In our view, forests, and the measures derived from them, indicate whether there is any predictive potential in a set of variables not yet integrated into a hypothesized SEM. In case these variables have not yet been considered in the theory that led to the specification of the statistical model, their discovery may lead researchers to reconsider and augment

their theory. In this endeavor, variable importance provides researchers with hints at which variables to select. How to structurally integrate these variables into the original SEM remains an open question. We hope that this article along with the freely provided SEM tree and forest software spawns future research to address this question in greater detail and in large data sets that contain formerly hidden and theoretically relevant predictive information.

SEM forests provide a structured approach to quantify variable importance in SEMs and support researchers in hypothesis generation and testing. As an extension of SEM trees, forests combine aspects of data-driven and theory-guided analysis in a single framework of theory-guided exploration (for an alternative, model-free approach, see Miller, Lubke, McArtor, & Bergeman, 2016). The cautionary note of Brandmaier et al. (2013b) that was expressed in the context of SEM trees, is just as valid for forests: Exploratory methods do not provide a shortcut from data to theories, nor from data to knowledge. Researchers are still required to think about their observations, remind themselves of the assumptions of their models, of the intricacies of the data sampling process, and the focus of their field; then, backed by this knowledge, make reasonable and responsible decisions about their analyses. Last but not least, a data-driven model must be evaluated on an independent dataset before its tenability can be claimed. If confirmatory analyses are planned, researchers are advised to conduct them before proceeding to exploratory approaches (McArdle, 2013; Tukey, 1962).

When to Grow More Than One Tree

When should we use trees and forests at all? When analyzing large data sets, scientists often would like to know which subset of variables has an influence on the phenomenon of interest (be this a univariate outcome, a correlation, a hypothesized causal relation, or any other relationship between or properties of variables) and, thus, carries the potential to explain the phenomenon. This exploratory approach is legitimate if the original hypotheses guiding the study turn out to be untenable, if the data set includes variables beyond the scope of extant theories, or whenever prediction does not only serve to validate a theory but actually is a goal in itself (e.g., in clinical decision making). The utility of complementing theory testing with tree-type exploration increases further when the number of predictors is large relative to the number of cases, and when effects of predictors on outcomes are interactive and nonlinear. If these conditions are met, the question remains when we should be content with a single tree and when we should grow more than one tree. We believe that in most cases both should be done as complementary analyses. First, a single tree may be grown to show a partitional structure of the sample inducing groups with different observed data patterns. The decision nodes of the tree may inform researchers about variables that describe differences and that may be better accounted for in future models. The partitional structure itself may be informative about the kind of differences between subgroups, for example, different factor profiles in factor analysis, or different growth curves in longitudinal data. However, it must be stressed that the resulting partition may neither be true nor be the best possible. It merely is one way of partitioning the sample that is optimal according to the chosen tree induction algorithm. Thus, it seems reasonable to turn to SEM forests in a subsequent analysis step. The forest analysis induces random variation to the empirical sample to obtain better estimates of variables' importance for predicting differences in observed data patterns. This comes at the cost of losing a straightforward way to recover a concrete partition from a forest.

Conclusion

In psychological research, the number of cases is often small compared with the number of variables measured but there are increasing number of studies that generate large data sets, for example, by using affordable and easily accessible online surveys, fused data sets from multisite studies, or data with a high density of measured variables. The latter is especially true for studies involving neuroimaging techniques or genetic associations. Likewise, purely behavioral data sets can benefit from exploratory analyses with trees and forests (e.g., if they comprise a large array of scales from various questionnaires). Tree and forest analyses may inform researchers about variables that provide additional information about the phenomenon or process they are hypothesizing about. Recent tree-based analyses of psychological data sets were conducted with longitudinally modeled child development (Brandmaier et al., 2013b), adult development (Brandmaier et al., 2013a), and late-life terminal decline (Ghisletta, 2013) of cognitive functioning across age, or perceptions of stress (Scott, Whitehead, Bergernan, & Pitzer, 2013). We share the hope of Strobl, Malley, and Tutz (2009) that trees will become a standard tool of analysis in psychological research and other empirical fields. The flexibility of structural equation modeling to account for many research designs and the flexibility of trees and forests to account for the diversity of predictors encountered in many settings makes the method suitable as a generic tool of exploration after a first step of purely theory-driven modeling. SEM trees and forests will enable researchers to make more efficient use of their empirical data. Lastly, trees may be one more step toward bringing individual differences back into the focus of psychological research.

References

Bender, A. R., Daugherty, A. M., & Raz, N. (2013). Vascular risk moderates associations between hippocampal subfield volumes and memory. *Journal of Cognitive Neuroscience, 25,* 1851–1862. http://dx.doi.org/10.1162/jocn_a_00435

Berk, R. (2006). An introduction to ensemble methods for data analysis. *Sociological Methods & Research, 34,* 263–295. http://dx.doi.org/10.1177/0049124105283119

Bertram, L., Böckenhoff, A., Demuth, I., Düzel, S., Eckardt, R., Li, S.-C., . . . Steinhagen-Thiessen, E. (2014). Cohort profile: The Berlin Aging Study II (BASE-II). *International Journal of Epidemiology, 43,* 703–712. http://dx.doi.org/10.1093/ije/dyt018

Böckenhoff, A., Sassenroth, D., Kroh, M., Siedler, T., Eibich, P., & Wagner, G. G. (2013). *The socio-economic module of the Berlin Aging Study II (SOEP-BASE): Description, structure, and questionnaire*. (SOEPpapers on Multidisciplinary Panel Data Research No. 568). Berlin, Germany: DIW.

Boker, S. M., Neale, M., Maes, H., Wilde, M., Spiegel, M., Brick, T. R., . . . Fox, J. (2011). OpenMx: An open source extended structural equation modeling framework. *Psychometrika, 76,* 306–317. http://dx.doi.org/10.1007/s11336-010-9200-6

Boker, S. M., 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. http://dx.doi.org/10.1007/978-1-4020-1958-6_9

Borsboom, D., Kievit, R. A., Cervone, D., & Hood, S. B. (2009). The two disciplines of scientific psychology, or: The disunity of psychology as a working hypothesis. In *Dynamic process methodology in the social and developmental sciences* (pp. 67–97). New York, NY: Springer. http://dx.doi.org/10.1007/978-0-387-95922-1_4

Brandmaier, A. M. (2015). *semtree: Recursive partitioning of structural equation models in R* [Computer software manual]. Retrieved from http://www.brandmaier.de/semtree

Brandmaier, A. M., von Oertzen, T., McArdle, J. J., & Lindenberger, U. (2013a). Exploratory data mining with structural equation model trees. In J. J. McArdle & G. Ritschard (Eds.), *Contemporary issues in exploratory data mining in the behavioral sciences* (pp. 96–127). New York, NY: Routledge.

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

Breiman, L. (1996). Bagging predictors. *Machine learning, 24,* 123–140. http://dx.doi.org/10.1023/A:1018054314350

Breiman, L. (2001a). Random forests. *Machine Learning, 45,* 5–32. http://dx.doi.org/10.1023/A:1010933404324

Breiman, L. (2001b). Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical Science, 16,* 199–231. http://dx.doi.org/10.1214/ss/1009213726

Breiman, L., & Cutler, A. (2014). *Random forests* [Computer software manual]. Retrieved from http://www.stat.berkeley.edu/breiman/RandomForests/cc_home.htm

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

Bühlmann, P., & Yu, B. (2002). Analyzing bagging. *The Annals of Statistics, 30,* 927–961. http://dx.doi.org/10.1214/aos/1031689014

Díaz-Uriarte, R., & De Andres, S. A. (2006). Gene selection and classification of microarray data using random forest. *BMC Bioinformatics, 7,* 3. http://dx.doi.org/10.1186/1471-2105-7-3

Düzel, S., Voelkle, M. C., Düzel, E., Gerstorf, D., Drewelies, J., Steinhagen-Thiessen, E., . . . Lindenberger, U. (2016). The subjective health horizon questionnaire (SHH-Q): assessing future time perspectives for facets of an active lifestyle. *Gerontology, 62,* 345–353. http://dx.doi.org/10.1159/000441493

Fernández-Delgado, M., Cernadas, E., Barro, S., & Amorim, D. (2014). Do we need hundreds of classifiers to solve real world classification problems? *Journal of Machine Learning Research, 15,* 3133–3181.

Ghisletta, P. (2013). Recursive partitioning to study terminal decline in the berlin aging study. In J. J. McArdle & G. Ritschard (Eds.), *Contemporary issues in exploratory data mining in the behavioral sciences* (pp. 405–428). New York, NY: Routledge.

Gower, J. C. (1966). Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika, 53,* 325–338. http://dx.doi.org/10.2307/2333639

Hapfelmeier, A., Hothorn, T., Ulm, K., & Strobl, C. (2014). A new variable importance measure for random forests with missing data. *Statistics and Computing, 24,* 21–34. http://dx.doi.org/10.1007/s11222-012-9349-1

Hapfelmeier, A., & Ulm, K. (2013). A new variable selection approach using random forests. *Computational Statistics & Data Analysis, 60,* 50–69. http://dx.doi.org/10.1016/j.csda.2012.09.020

Hastie, T., Tibshirani, R., & Friedman, J. (2001). *The elements of statistical learning*. New York, NY: Springer. http://dx.doi.org/10.1007/978-0-387-21606-5

Horn, J. L., & McArdle, J. J. (1992). A practical and theoretical guide to measurement invariance in aging research. *Experimental Aging Research, 18,* 117–144. http://dx.doi.org/10.1080/03610739208253916

Jöreskog, K. G. (1969). A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika, 34,* 183–202. http://dx.doi.org/10.1007/BF02289343

Jöreskog, K. G. (1970). Estimation and testing of simplex models. *British Journal of Mathematical and Statistical Psychology, 23,* 121–145. http://dx.doi.org/10.1111/j.2044-8317.1970.tb00439.x

Josefsson, M., de Luna, X., Pudas, S., Nilsson, L.-G., & Nyberg, L. (2012). Genetic and lifestyle predictors of 15-year longitudinal change in episodic memory. *Journal of the American Geriatrics Society, 60,* 2308–2312. http://dx.doi.org/10.1111/jgs.12000

Kaufman, A. S., Reynolds, C. R., & McLean, J. E. (1989). Age and WAIS-R intelligence in a national sample of adults in the 20- to 74-year age range: A cross-sectional analysis with educational level controlled. *Intelligence, 13,* 235–253. http://dx.doi.org/10.1016/0160-2896(89)90020-2

Kievit, R. A., Frankenhuis, W. E., Waldorp, L. J., & Borsboom, D. (2013). Simpson's paradox in psychological science: A practical guide. *Frontiers in Psychology, 4*. http://dx.doi.org/10.3389/fpsyg.2013.00513

Lamiell, J. T. (1981). Toward an idiothetic psychology of personality. *American Psychologist, 36,* 276–289. http://dx.doi.org/10.1037/0003-066X.36.3.276

Lang, F. R., John, D., Lüdtke, O., Schupp, J., & Wagner, G. G. (2011). Short assessment of the big five: Robust across survey methods except telephone interviewing. *Behavior Research Methods, 43,* 548–567.

Little, T. D., Lindenberger, U., & Nesselroade, J. R. (1999). On selecting indicators for multivariate measurement and modeling with latent variables: When "good" indicators are bad and "bad" indicators are good. *Psychological Methods, 4,* 192–211. http://dx.doi.org/10.1037/1082-989X.4.2.192

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

McArdle, J. J. (1994). Structural factor analysis experiments with incomplete data. *Multivariate Behavioral Research, 29,* 409–454. http://dx.doi.org/10.1207/s15327906mbr2904_5

McArdle, J. J. (2009). Latent variable modeling of differences and changes with longitudinal data. *Annual Review of Psychology, 60,* 577–605. http://dx.doi.org/10.1146/annurev.psych.60.110707.163612

McArdle, J. J. (2013). Exploratory data mining using decision trees in the behavioral sciences. In J. J. McArdle & G. Ritschard (Eds.), *Contemporary issues in exploratory data mining in the behavioral sciences* (pp. 3–47). New York, NY: Routledge.

McArdle, J. J., & Epstein, D. (1987). Latent growth curves within developmental structural equation models. *Child Development, 58,* 110–133. http://dx.doi.org/10.2307/1130295

McArdle, J. J., & Hamagami, F. (1992). Modeling incomplete longitudinal and cross-sectional data using latent growth structural models. *Experimental Aging Research, 18,* 145–166. http://dx.doi.org/10.1080/03610739208253917

McArdle, J. 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* (pp. 139–175). Washington, DC: American Psychological Association. http://dx.doi.org/10.1037/10409-005

McArdle, J. J., & Nesselroade, J. R. (2014). *Longitudinal data analysis using structural equation models*. Washington, DC: American Psychological Association.

McArdle, J. J., & Prescott, C. A. (1992). Age-based construct validation using structural equation modeling. *Experimental Aging Research, 18,* 87–115. http://dx.doi.org/10.1080/03610739208253915

Miller, P. J., Lubke, G. H., McArtor, D. B., & Bergeman, C. S. (2016). Finding structure in data using multivariate tree boosting. *Psychological Methods, 21,* 583–602.

Molenaar, P. C. M. (2004). A manifesto on psychology as idiographic science: Bringing the person back into scientific psychology, this time forever. *Measurement, 2,* 201–218. http://dx.doi.org/10.1207/s15366359mea0204_1

Muthén, B. O. (2004). Latent variable analysis: Growth mixture modeling and related techniques for longitudinal data. In D. Kaplan (Ed.), *The SAGE handbook of quantitative methodology for the social sciences* (pp. 345–370). Thousand Oaks, CA: SAGE Publications.

Nesselroade, J. R., Gerstorf, D., Hardy, S. A., & Ram, N. (2007). Focus article: Idiographic filters for psychological constructs. *Measurement: Interdisciplinary Research & Perspective, 5,* 217–235. http://dx.doi.org/10.1080/15366360701741807

Nesselroade, J. R., & Molenaar, P. C. M. (1999). Pooling lagged covariance structures based on short, multivariate time series for dynamic factor analysis. In R. H. Hoyle (Ed.), *Statistical strategies for small sample research Nesselroade* (pp. 223–251). Newbury Park, CA: Sage.

Nordahl, C. W., Ranganath, C., Yonelinas, A. P., DeCarli, C., Reed, B. R., & Jagust, W. J. (2005). Different mechanisms of episodic memory failure in mild cognitive impairment. *Neuropsychologia, 43,* 1688–1697. http://dx.doi.org/10.1016/j.neuropsychologia.2005.01.003

Quinlan, J. (1986). Induction of decision trees. *Machine Learning, 1,* 81–106. http://dx.doi.org/10.1007/BF00116251

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

Raz, N., Rodrigue, K. M., & Acker, J. D. (2003). Hypertension and the brain: Vulnerability of the prefrontal regions and executive functions. *Behavioral Neuroscience, 117,* 1169–1180. http://dx.doi.org/10.1037/0735-7044.117.6.1169

R Core Team. (2013). *R: A language and environment for statistical computing* [Computer software manual]. Vienna, Austria. Retrieved from http://www.R-project.org/

Reynolds, C. R., Chastain, R. L., Kaufman, A. S., & McLean, J. E. (1987). Demographic characteristics and IQ among adults: Analysis of the WAIS-R standardization sample as a function of the stratification variables. *Journal of School Psychology, 25,* 323–342. http://dx.doi.org/10.1016/0022-4405(87)90035-5

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics, 20,* 53–65. http://dx.doi.org/10.1016/0377-0427(87)90125-7

Schulz, P., & Schlotz, W. (1999). The trier inventory for the assessment of chronic stress (tics): scale construction, statistical testing, and validation of the scale work overload. *Diagnostica, 45,* 8–19.

Scott, S. B., Whitehead, B. R., Bergernan, C. S., & Pitzer, L. (2013). Understanding global perceptions of stress in adulthood through tree-based exploratory data mining. In J. Mcardle & G. Ritschard (Eds.), *contemporary issues in exploratory data mining in the behavioral sciences* (pp. 371–404). New York, NY: Routledge.

Shing, Y. L., Werkle-Bergner, M., Brehmer, Y., Müller, V., Li, S. C., & Lindenberger, U. (2010). Episodic memory across the lifespan: The contributions of associative and strategic components. *Neuroscience & Biobehavioral Reviews, 34,* 1080–1091. http://dx.doi.org/10.1016/j.neubiorev.2009.11.002

Shmueli, G. (2010). To explain or to predict? *Statistical Science, 25,* 289–310. http://dx.doi.org/10.1214/10-STS330

Simpson, E. H. (1951). The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society: Series B, 13,* 238–241.

Singer, T., Verhaeghen, P., Ghisletta, P., Lindenberger, U., & Baltes, P. B. (2003). The fate of cognition in very old age: Six-year longitudinal findings in the berlin aging study (base). *Psychology and Aging, 18,* 318–331. http://dx.doi.org/10.1037/0882-7974.18.2.318

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.-L., Kneib, T., Augustin, T., & Zeileis, A. (2008). Conditional variable importance for random forests. *BMC Bioinformatics, 9,* 307. http://dx.doi.org/10.1186/1471-2105-9-307

Strobl, C., Boulesteix, A., Zeileis, A., & Hothorn, T. (2007). Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics, 8,* 25. http://dx.doi.org/10.1186/1471-2105-8-2

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,* 323–348. http://dx.doi.org/10.1037/a0016973

Strobl, C., & Zeileis, A. (2008). *Danger: High power! Exploring the statistical properties of a test for random forest variable importance* (Tech. Rep. No. 017). Department of Statistics, University of Munich.

Tucker, L. (1966). Learning theory and multivariate experiment: Illustration by determination of generalized learning curves. In R. B. Cattell (Ed.), *Handbook of multivariate experimental psychology* (pp. 476–501). Chicago. IL: Rand McNally & Company.

Tukey, J. W. (1962). The future of data analysis. *The Annals of Mathematical Statistics, 33,* 1–67. http://dx.doi.org/10.1214/aoms/1177704711

Verikas, A., Gelzinis, A., & Bacauskiene, M. (2011). Mining data with random forests: A survey and results of new tests. *Pattern Recognition, 44,* 330–349. http://dx.doi.org/10.1016/j.patcog.2010.08.011

Vermunt, J. K., & Magidson, J. (2002). Latent class cluster analysis. In J. Hagenaars & A. McCutcheon (Eds.), *Applied latent class analysis* (pp. 89–106). Cambridge, UK: Cambridge University Press.

VonDras, D. D., Powless, M. R., Olson, A. K., Wheeler, D., & Snudden, A. L. (2005). Differential effects of everyday stress on the episodic memory test performances of young, mid-life, and older adults. *Aging & Mental Health, 9,* 60–70. http://dx.doi.org/10.1080/13607860412331323782

von Oertzen, T., Brandmaier, A. M., & Tsang, S. (2015). Structural equation modeling with Ωnyx. *Structural Equation Modeling: A Multidisciplinary Journal, 22,* 148–161. http://dx.doi.org/10.1080/10705511.2014.935842

Watson, D., Clark, L. A., & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: The PANAS scales. *Journal of personality and social psychology, 54,* 1063–1070. http://dx.doi.org/10.1037/0022-3514.54.6.1063

Wolf, O. T. (2009). Stress and memory in humans: Twelve years of progress? *Brain Research, 1293,* 142–154.

Appendix A

Descriptive Statistics of Simulated Data for a Factor Model Forest

Variable	x1	x2	x3	cov1	cov2	cov3	cov4	cov5	cov6
N	500.00	500.00	500.00	500.00	500.00	500.00	500.00	500.00	500.00
Mean	1.05	1.03	1.08	.50	.51	.52	.49	.48	.51
SD	2.76	2.62	2.71	.50	.50	.50	.50	.50	.50
x1	1.00	.87	.87	.74	.72	−.22	−.22	.02	.07
x2	.87	1.00	.87	.74	.73	−.23	−.17	.02	.08
x3	.87	.87	1.00	.74	.73	−.20	−.20	−.02	.12
cov1	.74	.74	.74	1.00	.76	.02	.01	−.01	.04
cov2	.72	.73	.73	.76	1.00	.02	.00	.02	.08
cov3	−.22	−.23	−.20	.02	.02	1.00	.19	−.03	.00
cov4	−.22	−.17	−.20	.01	.00	.19	1.00	−.04	.02
cov5	.02	.02	−.02	−.01	.02	−.03	−.04	1.00	.01
cov6	.07	.08	.12	.04	.08	.00	.02	.01	1.00

Note. Observed variables in the factor model are x1, x2, and x3. Predictors are cov1 to cov6.

Appendix B

Example of a Factor Model Forest

To generate a SEM forest and visualize the results only a few steps are needed. In general, an OpenMx model, a dataset with variables observed in the model and covariates included, and a set of control parameters are the only requirements. We will provide steps for fitting an OpenMx model with data, creating a forest, and plotting variance importance.

The OpenMx Model

SEM trees require model specification in OpenMx (Boker et al., 2011), a package for the statistical programming language R (R Core Team, 2013). OpenMx allows path and matrix specification of models. In the following, we present a path specification of a factor model with three observed variables (x1–x3) that are assumed to measure a single construct. The observed variables are indicated by a single latent factor, with all shared variance contained at the latent level, and unique variation residing in the residuals for each indicator. The following code will create a factor SEM, fit the dataset to the model, and provide model parameters and fit indices to the console.

(*Appendices continue*)

```
data(factorData)
factorModel <- mxModel(
  "Factor Model",
  type="RAM", mxData(factorData,type="raw"),
  manifestVars = c("x1","x2","x3"),
  latentVars = "f",
  mxPath(from="f",to=c("x1","x2","x3"),
                free=c(FALSE,TRUE,TRUE),
                value=c(1,1,1), arrows=1, label=c("l1","l2","l3")),
  mxPath(from="one",to=c("f"), free=c(TRUE),
                value=c(1.0), arrows=1, label=c("mu_f")),
  mxPath(from="f",to=c("f"), free=c(TRUE),
                value=c(1.0), arrows=2, label=c("var_f")),
  mxPath(from="x1",to=c("x1"), free=c(TRUE),
                value=c(1.0), arrows=2, label=c("e")),
  mxPath(from="x2",to=c("x2"), free=c(TRUE),
                value=c(1.0), arrows=2, label=c("e")),
  mxPath(from="x3",to=c("x3"), free=c(TRUE),
                value=c(1.0), arrows=2, label=c("e"))
) # close model
summary(factorFit <- mxRun(factorModel))
```

Starting estimation with the full dataset using `mxRun()` will indicate errors in model syntax, or potential model estimation issues, prior to forest creation with the command `semforest()`. The following example uses a simulated dataset as presented in Appendix A. Three indicators are shown together with six covariates. Two covariates have high intercorrelations, two variables have moderate intercorrelation, and two variables are uncorrelated random noise. Each covariate is coded as two categories (0/1), with standard normal effects of cov1 and cov2 distributed as $\mathcal{N}(2.0, 0.3)$ and cov3 and cov4 distributed as $\mathcal{N}(1.0, 0.5)$ on the latent level factor score. The intention of this data structure is to show the utility of the SEM forest method when compared with a simple SEM tree analysis. When unmodeled covariates have overlapping information to different degrees, their impact within the model may be lost.

Creating a Control Object

Users have control over parameters determining the growth process of a forest. The parameters may be changed depending on the model, the number of covariates, and the research hypotheses to be tested. The forest control options contain:

- `num.trees` - how many trees to create in the SEM Forest process.
- `sampling` - method for selecting cases in each SEM Tree.
- `control` - a SEM Tree control object as described in Brandmaier, von Oertzen, McArdle, & Lindenberger (2013b).
- `mtry` - number of covariates to test at each node for splitting algorithm.

A control object with default settings is done with the following line:

```
factorControls <- semforest.control()
# Change the Default settings in semforest.control() and semtree.control()
factorControls$num.trees <- 1000
factorControls$semtree.control$method <- "naive"
```

Control objects will shape the forest (and individual trees) to fall within parameters established by the researcher. Please note the defaults are supplied not as suggested starting points, but in order to prevent computational overloads for novice users.

SEM Tree Interpretation

A SEM tree is fit to the factor model structure with the simulated dataset. Using the control object from the above step, we run the `semtree()` function:

```
factorTree <- semtree(model=factorModel, data=factorData, semforest.control=
  factorControls$semtree.control)
```

(*Appendices continue*)

The individual tree is the standard "naïve" splitting process, where all split points have the same potential influence independent of which covariate they indicate. Plotting a tree output is relatively straightforward, with multiple tree complications encountered in a random forest analysis.

```
plot(factorTree)
```

The output of this plot is shown in Figure B1. Of note is the structure of the splits found for subsetting the data at each node. Cov1 is selected initially (but was equally as influential as cov2 in the simulation design), while cov3 and cov2 were selected for secondary splits (with the impact of cov4 equal to cov3).

Creating a Forest

The semforest command uses the elements described above to compile SEM trees following random forest logic. The most basic tree can be run by the following command line:

```
factorForest <- semforest(model=factorModel, data=factorData, Semforest.
  control=factorControls)
```

Alternatively, single trees can be extracted from a forest as follows:

```
tree42 <- factorForest$forest[[42]]  # Retrieve the 42nd tree from the
  forest
```

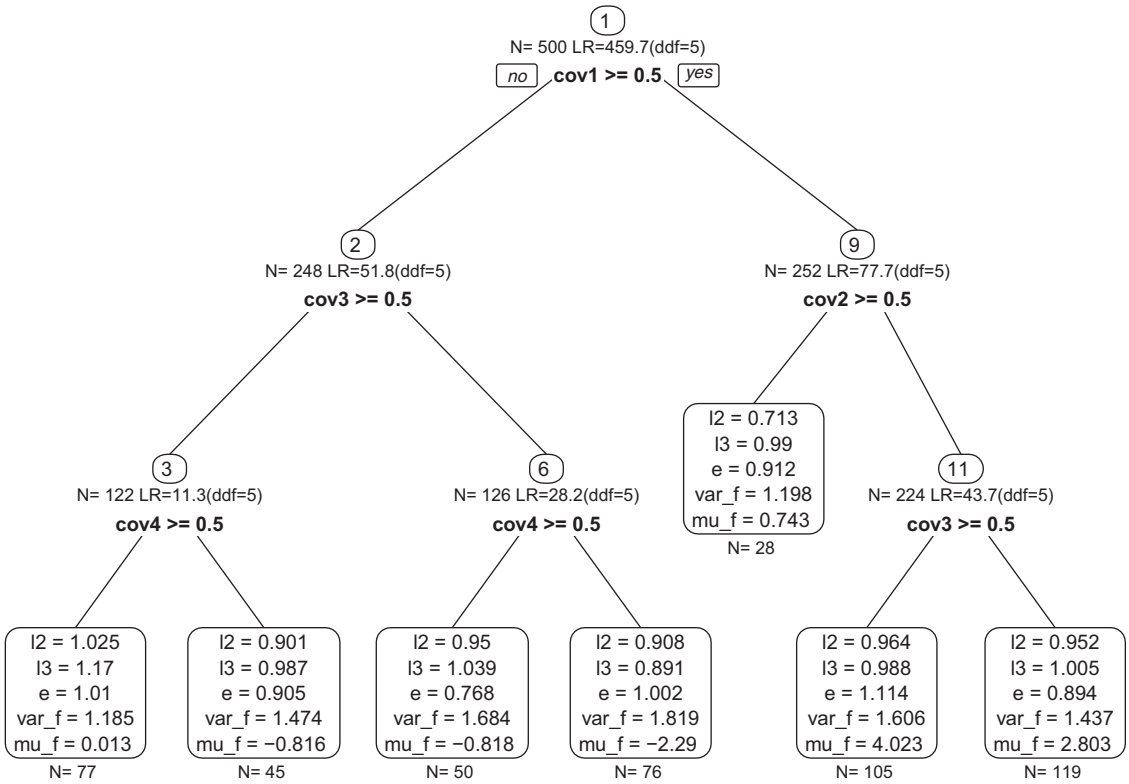

Figure B1. Tree output for simulated data with six covariates. Covariates 1 and 2 had the highest impact, Covariates 3 and 4 had low impacts, and Covariates 5 and 6 represent random noise variables.

(*Appendices continue*)

The semforest() arguments are in the following order: SEM OpenMx model, dataset (with model variables and covariates: cov1-cov6), and researcher specified control object. This step may take some time to complete. On machines with more than one core available (and in cases where grid computing is available), users may want to use parallel processing to speed up computation of *n* trees. The R package "snowfall" is required and may be implemented in the following way:

```
require(snowfall)
sfInit(parallel=TRUE, cpus=2)
sfClusterEval(require("OpenMx"))
sfClusterEval(require("semtree"))
```

The above code will initialize a cluster of two CPUs in parallel and load the required packages remotely. The command `semforest()` will detect this environment automatically to distribute jobs over the available CPUs.

Visualizing a Semforest Object

Once a semforest has been computed, a couple of tools are available to researchers to print and visualize results. Variable Importance is a graphical display of the impact of each potential splitting covariate for subsetting the overall model into a two group nested model. The impact of each covariate is graphed as bar plots or as average model improvement traced over iterations.

```
forestVarimp <- varimp(factorForest, parallel=FALSE)
plot(forestVarimp) # Average improvement bar plot
```

The plot shown in Figure B2 is the output of this variable importance plot function. The x-axis shows the average model improvement ($\Delta - 2LL$) between the overall model and the two group model based on subsetting the data by splitting on a covariate. The top are the variables that improve model fit the most and the bottom variables improve the model the least. Negative values indicate that the splits provide no information about splitting the sample. This figure shows concisely that cov1 and cov2 have about equal importance for the forest analysis, cov3 and cov4 are also about equal in their influence, and cov5 and cov6 both have no influence on the forest model. This outcome is at odds with the tree shown in Figure B1, because cov1 and cov2 are shown to be equally important, where a single tree indicates that cov1 is the most important and cov2 only influences the right branch.

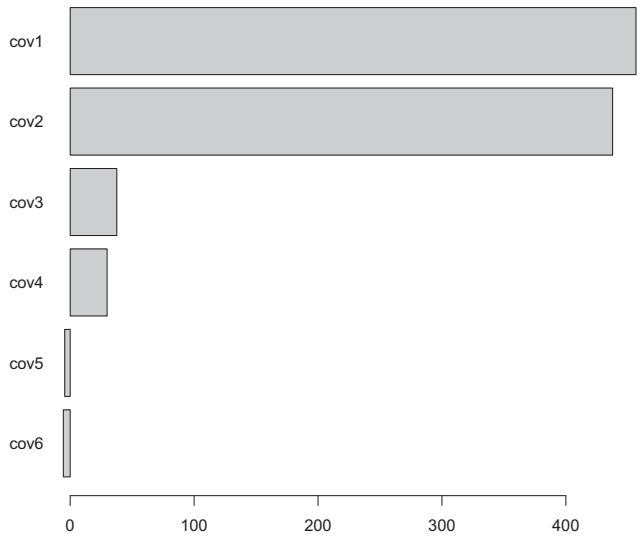

Figure B2. Variable importance plot for the simulated data fit to a factor model. Cov1 and cov2 have about equal importance for the forest analysis, cov3 and cov4 are also about equal in their influence, and cov5 and cov6 both have no influence on the forest model.

Received June 1, 2015
Revision received March 7, 2016
Accepted April 3, 2016 ■