Network Trees: A Method for Recursively Partitioning Covariance Structures

Payton J. Jones
Harvard University

Patrick Mair
Harvard University

Thorsten Simon
Universität Innsbruck

Achim Zeileis
Universität Innsbruck

Abstract

In many areas of psychology, correlation-based network approaches (i.e., psychometric networks) have become a popular tool. In this paper we propose an approach that recursively splits the sample based on covariates in order to detect significant differences in the structure of the covariance or correlation matrix. Psychometric networks or other correlation-based models (e.g., factor models) can be subsequently estimated from the resultant splits. We adapt model-based recursive partitioning and conditional inference tree approaches for finding covariate splits in a recursive manner. The empirical power of these approaches is studied in several simulation conditions. Examples are given using real-life data from personality and clinical research.

Keywords: network analysis, correlation networks, recursive partitioning, decision trees, conditional inference.

1. Introduction

Network science is part of the larger field of complex systems, which studies how associations between many parts in a system give rise to emergent properties. Within the past decade, the network approach has gained significant ground as an approach for studying mental disorders as systems of interacting symptoms or other psychobiological mechanisms (McNally 2019; Borsboom 2017; Jones, Heeren, and McNally 2017). The rapid success of network theory can perhaps be attributed to a historically uncomfortable mismatch between categorical disease models of psychopathology and cognitive-behavioral models. It has been argued that diagnostic manuals (e.g., American Psychiatric Association 2013) implicitly follow a common-cause model that depicts mental disorders as categorical disease entities that give rise to emotional and behavioral symptoms (Borsboom 2017). In contrast, cognitive-behavioral models emphasize the importance of self-reinforcing feedback loops among transdiagnostic behaviors, emotions, and cognitions, such as the pathological role of avoidance behaviors in reinforcing anxiety (Hanley, Iwata, and McCord 2003; Salters-Pedneault, Tull, and Roemer 2004). Such relationships are often disallowed or disregarded in common-cause models that view mental disorders as underlying diseases (Borsboom 2017).

In most cases, relationships between psychological variables cannot be observed directly. Instead, these relationships must be estimated based on observed properties, necessitating the development of statistical techniques to apply network analytic methods in psychology. These methods have been broadly dubbed *network psychometrics*, and typically involve estimating conditional dependence graphs from covariance matrices (Epskamp, Rhemtulla, and Bors-

This is a preprint of an article published in *Psychometrika*, **85**(4), 926–945.
doi:10.1007/s11336-020-09731-4

boom 2017; Epskamp, Borsboom, and Fried 2018; Marsman, Borsboom, Kruis, Epskamp, Van Bork, Waldorp, Van der Maas, and Maris 2018). In addition to their role in studying mental disorders, network psychometric approaches have been used to study personality (Costantini, Richetin, Preti, Casini, Epskamp, and Perugini 2019), attitudes (Dalege, Borsboom, Van Harreveld, Van den Berg, Conner, and Van der Maas 2016), resilience factors (Fritz, Fried, Goodyer, Wilkinson, and Van Harmelen 2018), and other psychological systems.

Unfortunately, common network psychometric approaches are limited by assumptions of homogeneity. The systems that constitute mental disorders and other psychological phenomena are argued to be extremely heterogeneous (Fried and Nesse 2015). In cases where network models do not account for potential moderating factors, networks may blur important distinctions in causal mechanisms that are heterogeneous across different groups. Indeed, in some cases averaged models may not be representative of any given subset and effects may even be sign reversed (Boker and Martin 2018; Molenaar 2004). It is possible that networks may vary across gender, age, personality, and a host of additional variables. There is thus a natural marriage between network models and statistical approaches that allow for explicitly modeling heterogeneity using appropriate covariates.

There are some existing approaches to modelling network heterogeneity using covariates. Perhaps most relevant are moderated network models (MNMs). MNMs allow each pairwise associations between variables in a network to be moderated by covariates (Haslbeck, Borsboom, and Waldorp 2019). MNMs are an approach for exploring covariates in networks in a parameterized manner. Another approach is to separate subgroups that are a priori hypothesized to differ, estimate the networks separately, and compare them. Comparisons can be evaluated through permutation testing, often referred to as a Network Comparison Test (Van Borkulo, Boschloo, Kossakowski, Tio, Schoevers, Borsboom, and Waldorp 2017).

Changepoint analyses are an example of incorporating a very specific covariate to networks: time. Time is unique compared to other covariates because it introduces dependency between each sequential observation. Various approaches to changepoint analysis have been discussed in the network psychometric literature, including kernel-based and hidden Markov change point models (Cabrieto, Tuerlinckx, Kuppens, Wilhelm, Liedlgruber, and Ceulemans 2018; Park and Sohn 2019).

In this article we propose an approach that recursively splits the sample based on covariates in order to detect differences in the covariance (or correlation) matrix, which forms the basis of parameters in correlation-based network approaches. The goal of this approach is to identify meaningful subgroups in the data. Such recursive partitioning approaches have a long tradition in the statistical literature and became increasingly popular in psychology in recent years (e.g., Strobl, Malley, and Tutz 2009; Merkle and Shaffer 2011; Brandmaier, von Oertzen, McArdle, and Lindenberger 2013). One benefit of this approach compared to MNMs is that it can automatically detect nonlinearities and interactions in an exploratory setting (for details see Zeileis, Hothorn, and Hornik 2008; Hothorn, Hornik, and Zeileis 2006b). Whereas MNMs are particularly useful for exploring moderation of individual pairwise associations, recursive partitioning facilitates interpretations of how covariates are related to heterogeneity across the entire network.

An important development within this context is *model-based recursive partitioning* (MOB; Zeileis *et al.* 2008). It is a semi-parametric approach which aims to finds splits with respect to parameters of a particular underlying model. In other words, we have a formal parametric model as "response" in each terminal node, rather than a single dependent variable only as in classical tree approaches like CART (classification and regression trees; Breiman, Friedman, Olshen, and Stone 1984). Using the MOB algorithm the splits are determined in a way such that the model parameters are maximally heterogeneous across the terminal nodes. In the psychometric literature the MOB idea has been adapted to several model families: Bradley-Terry models (Strobl, Wickelmaier, and Zeileis 2011), Rasch models (Strobl, Kopf, and Zeileis 2015; Komboz, Strobl, and Zeileis 2018), factor analysis and structural equation models (Merkle and Zeileis 2013), multinomial processing trees (Wickelmaier and Zeileis 2018), and generalized linear mixed-effects models (Fokkema, Smits, Zeileis, Hothorn, and Kelderman 2018).

Another recursive partitioning approach proposed in the statistical literature are *conditional inference trees* (CTree; Hothorn *et al.* 2006b). CTree is very similar to MOB in many respects but does not have to be based on a formal parametric model. Instead, CTree is based on a general class of permutation tests which can be combined with maximum likelihood scores, but can also be based on completely nonparametric test statistics. While originally introduced for simple decision trees with mean responses in the terminal nodes, CTree has also recently been leveraged to obtain trees with fitted parametric models in each node (see, e.g., Seibold, Zeileis, and Hothorn 2016).

In this article we integrate both MOB and CTree into network models in order to find optimal subgroup splits which lead to heterogeneous covariance structures in the terminal nodes. We chose these two frameworks because of their current status as leading methods for unbiased recursive partitioning in parametric inference and nonparametric inference, respectively. We begin with technical elaborations on MOB and CTree within the context of network models, with the formulation of a multivariate normal model at its core. Similarities and differences between the two approaches are discussed. Next, some simulation studies are performed exploring sensitivity and specificity of MOB and CTree networks. This is followed by a real-life application. R (R Core Team 2020) code for the simulations and an R Markdown document fully reproducing the applied example are provided in the supplemental materials (`https://osf.io/ykq2a/`).

2. Network tree theory

This section formally introduces the network tree (NT) approach. We develop the NT theory for both MOB and CTree. The basic algorithm, inherent to both tree approaches, is the following:

1. fit a multivariate normal distribution to the response in the current sub-sample (starting with the full sample in the first step);
2. compute a goodness-of-fit measure for each data point with respect to the parameters of the multivariate normal distribution (e.g., based on the log-likelihood);
3. assess whether this measure is correlated with any of the partitioning variables (construction of test statistics and selecting a partitioning variable using a mininmum p-value strategy);
4. find the split that yields the highest improvement of the goodness-of-fit measure.

These steps are repeated until no further split would significantly increase the goodness-of-fit measure. Other stopping criteria such as the minimum sample size per node can be defined as well.

Even though MOB and CTree are conceptually similar, there are some differences in how MOB and CTree tackle these four steps. In Step 3, MOB uses parameter instability tests based on the asymptotic multivariate normal distribution of the maximum likelihood estimators, whereas CTree uses permutation tests of empirical correlations. In Step 4, MOB maximizes a partitioned log-likelihood, whereas CTree maximizes a two-sample statistic. A systematic comparison of these two approaches follows below, after a formal introduction of MOB and CTree networks.

2.1. MOB networks

Since the models considered in this paper are based on Pearson correlations, we assume an underlying multivariate normally distributed p-dimensional random variable Y with individual observations $\mathbf{y}_i$ ($i = 1, \dots, n$; collected as $n \times p$ matrix $\mathbf{Y}$). The corresponding density function with mean vector $\boldsymbol{\mu}$ and variance-covariance matrix $\boldsymbol{\Sigma}$ is

$$f(\mathbf{y}_i; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{\sqrt{(2\pi)^p |\boldsymbol{\Sigma}|}} \exp\left(-\frac{1}{2}(\mathbf{y}_i - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\mathbf{y}_i - \boldsymbol{\mu})\right). \tag{1}$$

Let $\mathbf{D}$ be the diagonal matrix with the reciprocals of the square roots of the diagonal entries in $\boldsymbol{\Sigma}$ on its diagonal. Then $\mathbf{R} = \mathbf{D}\boldsymbol{\Sigma}\mathbf{D}^\top$ is the $p \times p$ correlation matrix.

Let $\boldsymbol{\theta} = (\boldsymbol{\mu}, \boldsymbol{\sigma}^2, \boldsymbol{\rho})$ be the full parameter vector of dimension k containing p means in $\boldsymbol{\mu}$, p variances in $\boldsymbol{\sigma}^2$, and $p(p-1)/2$ correlations from the lower triangular part of $\mathbf{R}$, collected in $\boldsymbol{\rho}$. The associated log-likelihood is

$$\begin{aligned} \ell(\boldsymbol{\theta}; \mathbf{y}_1, \dots, \mathbf{y}_n) &= \sum_{i=1}^{n} \ell(\boldsymbol{\theta}, \mathbf{y}_i) = \sum_{i=1}^{n} \log f(\mathbf{y}_i; \boldsymbol{\theta}) \\ &= \sum_{i=1}^{n} -\frac{1}{2}\left((\mathbf{y}_i - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\mathbf{y}_i - \boldsymbol{\mu}) + \log|\boldsymbol{\Sigma}| + p\log(2\pi)\right). \end{aligned} \tag{2}$$

Taking the partial derivatives with respect to the parameters leads to the *score function* (see the Appendix for the details):

$$\mathbf{s}(\boldsymbol{\theta}; \mathbf{y}_i) = \left(\frac{\partial \ell(\boldsymbol{\theta}; \mathbf{y}_i)}{\partial \theta_1}, \dots, \frac{\partial \ell(\boldsymbol{\theta}; \mathbf{y}_i)}{\partial \theta_k}\right)^\top. \tag{3}$$

Solving $\sum_{i=1}^{n} \mathbf{s}(\hat{\boldsymbol{\theta}}; \mathbf{y}_i) = \mathbf{0}$ gives the maximum likelihood (ML) estimates $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\sigma}}^2, \hat{\boldsymbol{\rho}})$. Note that in the case of a multivariate normal distribution without further covariates no numeric optimization is required since there are closed form solutions for all the parameters in $\boldsymbol{\theta}$. In the unrestricted case, the sample mean and covariance matrix (divided by n) lead to the maximum likelihood estimates.

Let us elaborate on Eq. (3) in a more narrative way. The score function is of key importance in the MOB approach since it acts as a goodness-of-fit measure. Each individual i has k scores, one for each parameter in vector $\boldsymbol{\theta}$. The starting point is the assumption that the same model parameters $\boldsymbol{\theta}$ hold for all individuals. In this case, all individual scores should be close to 0 and randomly fluctuate around it, much like least squares residuals in a linear regression model. Thus, the $n \times k$ score matrix can be used as a measure how well each of the k estimated parameters fit to the n individual observations. If the constant parameter assumption does not hold we have a case of non-invariance. For instance, say we have two subgroups A and B for which the true parameters differ. At an individual level, this non-invariance is reflected in the scores as follows: scores of individuals in A are negative, and scores of individuals in B are positive. This fact gives evidence that the sample should be split with respect to this grouping variable. More generally, if the subgroups are not coded directly by such an A-B grouping variable but parameters change along some continuous variable (such as age), then the scores of the constant model fit switch from positive to negative (or vice versa) along that variable.

We now need construct a test statistic that allows us to test explicitly for such parameter invariance/non-invariance. We adopt a class of structural change tests that form the basis of the MOB framework. While originating in the econometrics literature for testing changes over time (*generalized M-fluctuation tests*; Zeileis and Hornik 2007), these tests have also been adopted for assessing parameter constancy along other variables in cross-sectional data (see Merkle and Zeileis 2013, for an adaptation to psychometric models). Note that for notation convenience we limit our elaborations to a single partitioning variable Z only. The extension to the practically relevant case of multiple variables $Z_1, \ldots, Z_q$ is straightforward, since the score process and subsequent test statistics, as introduced below, are computed for each of the q variables. The starting point for the invariance testing framework is the *cumulative score process*[1]

$$\mathbf{B}(t; \hat{\boldsymbol{\theta}}) = \hat{\mathbf{I}}^{-1/2} n^{-1/2} \sum_{i=1}^{\lfloor nt \rfloor} \mathbf{s}(\hat{\boldsymbol{\theta}}; \mathbf{y}_{(i)}) \tag{4}$$

which is computed for each of the q partitioning variables separately. Here, $\mathbf{y}_{(i)}$ denotes the ith ordered observation with respect to the partitioning variable Z. $\hat{\mathbf{I}} = \sum_{i=1}^{n} s(\hat{\boldsymbol{\theta}}; \mathbf{y}_i) s(\hat{\boldsymbol{\theta}}; \mathbf{y}_i)^\top$ is an estimate of the covariance matrix of the scores, and $\lfloor nt \rfloor$ is the floor function of nt with t as a fraction of the sample size. The observed values of t are restricted to the set $\{0, 1/n, 2/n, \ldots, n/n\}$ such that $\mathbf{B}(t; \hat{\boldsymbol{\theta}})$ becomes an $n \times k$ matrix with elements $\mathbf{B}(i/n; \hat{\boldsymbol{\theta}})_{ij}$. For simplicity in notation we denote this matrix by $\mathbf{B}(\hat{\boldsymbol{\theta}})$, with elements $\mathbf{B}(\hat{\boldsymbol{\theta}})_{ij}$.

Based on $\mathbf{B}(\hat{\boldsymbol{\theta}})$ various test statistics can be constructed. We limit our explanations to *(maximum) Lagrange multiplier* (LM) tests, as presented in Andrews (1993), Merkle and Zeileis (2013), Merkle, Fan, and Zeileis (2014), and Wang, Merkle, and Zeileis (2014), for which desirable power properties have been shown. Depending on the scale of Z, various test LM statistics can be constructed. For Z metric

$$\max LM = \max_{i = \underline{i}, \ldots, \bar{i}} \left\{ \frac{i}{n} \left(1 - \frac{i}{n} \right) \right\}^{-1} \sum_{j=1}^{k} \mathbf{B}(\hat{\boldsymbol{\theta}})_{ij}^2, \tag{5}$$

where $[\underline{i}, \bar{i}]$ denotes the interval for the potential change point shift ($\underline{i}$ as the minimum segment size, $\bar{i} = n - \underline{i}$).

In this statistic the maximum is taken across all possible divisions of individuals into two groups with respect to Z. Additionally, the statistic is scaled by the asymptotic variance $t(1-$

[1] For further details on this process and its asymptotic properties see Hjort and Koning (2002) and Zeileis and Hornik (2007).

t) of the process $B(t; \hat{\boldsymbol{\theta}})$ (Merkle *et al.* 2014). For $\max LM$, critical values and corresponding p-values can be found by approximate asymptotic techniques (Hansen 1997).

If Z is ordinal, ties are typically present and therefore there is only partial ordering of the individuals. Merkle *et al.* (2014) proposed the following modification of (5) which involves binning of individuals at each of the m levels of Z:

$$\max LM_o = \max_{i \in \{i_1, \dots, i_{m-1}\}} \left\{ \frac{i}{n} \left(1 - \frac{i}{n} \right) \right\}^{-1} \sum_{j=1}^{k} \mathbf{B}(\hat{\boldsymbol{\theta}})_{ij}^2. \tag{6}$$

Critical values and p-values can be found by direct simulation (Zeileis 2006).

In case of nominal Z (m levels) there is obviously no ordering. Hjort and Koning (2002) proposed an LM test statistic which first sums the scores within each of the m levels of Z, followed by aggregating these sums across all m levels:

$$LM_{uo} = \sum_{l=1}^{m} \sum_{j=1}^{k} \left(\mathbf{B}(\hat{\boldsymbol{\theta}})_{i_l j} - \mathbf{B}(\hat{\boldsymbol{\theta}})_{i_{l-1} j} \right)^2. \tag{7}$$

Critical values and associated p-values can be obtained through a χ^2-distribution, as this statistic is asymptotically equivalent to the usual likelihood-ratio statistic.

In the MOB algorithm, where all q partitioning variables are considered, a separate cumulative score process and subsequent test statistic is computed for each Z. The variable associated with the smallest significant p-value (Bonferroni corrected for the number of partitioning variables) is selected for binary splitting (i.e., minimum p-value strategy). The split sample in the daughter nodes is then subject to further potential splitting, using the same procedure. The algorithm stops when no further significant splits in the partitioning variables can be found. Eventually, the terminal nodes yield a non-overlapping partition of the sample.

An attractive feature of the MOB approach is that we can restrict invariance testing to a subset of the parameters in $\boldsymbol{\theta}$ ("partial structural changes"; Andrews 1993; Merkle and Zeileis 2013). For example, in our network approach, we may only be interested in exploring invariance in the correlation parameters $\boldsymbol{\rho}$ and not in the variances and means.

2.2. CTree networks

In its basic form, CTree finds predictor splits with respect to single or multiple response variables. As opposed to algorithms like CART, CTree uses statistical tests for splitting by embedding the recursive partitioning idea into the conditional inference permutation test framework proposed by Strasser and Weber (1999, see also Hothorn, Hornik, Van de Wiel, and Zeileis 2006a). This framework uses independence tests constructed by means of the conditional distribution of linear test statistics.

In the CTree approach we operate on standardized versions of the variables in $\mathbf{Y}$, since it is sufficient to derive a proportionality expression for the correlation vector involving a simple vector multiplication, as formalized below. Therefore, let $\mathbf{Y}^*$ be the $n \times p$ data matrix ($i = 1, \dots, n$; $j = 1, \dots, p$) of standardized variables. The row vector for participant i we denote by $\mathbf{y}^*_{i\bullet}$, whereas the column vector for variable j is $\mathbf{y}^*_{\bullet j}$. As in the MOB approach above, we limit our explanations to a single paritioning variable Z only, with z_i indicating a single observation. The generalization to multiple Z's is straightforward as the test below is carried out for each of these variables separately.

The linear test statistic used in CTree networks essentially captures the linear association between (a transformation of) the partitioning variable Z and (a transformation of) the standardized multivariate response $\mathbf{Y}^*$:

$$\mathbf{T} = \mathrm{vec}\left(\sum_{i=1}^{n^*} g(z_i) h(\mathbf{y}^*_{i\bullet})^\top\right). \tag{8}$$

The sum goes from i to n^*, with n^* denoting the sample size within a particular branch of the tree. Before the first split, $n^* := n$. The "vec" operator re-organizes the resulting matrix as column vector by stacking the columns on top of each other.

Let us elaborate in detail on the intuition behind the test statistic in Eq. (8). Let us assume for the moment that $g(\cdot)$ and $h(\cdot)$ are identity functions. What is left is the sum of the products of z_i and $\mathbf{y}^*_{i\bullet}$. These components are the building blocks of a Pearson correlation. Therefore, $\mathbf{T}$ leads to a correlation test. For computing the corresponding p-values the linear statistic $\mathbf{T}$ has to be standardized by its corresponding covariance matrix and, in the multivariate case, aggregated to a scalar statistic using either a sum of squares (default) or by taking the maximum.

For our NT framework we use the following specifications. The covariate transformation function is $g(\cdot)$. If Z is metric or ordinal, the identity function is used and $g(z_i)$ is scalar. In case of nominal Z, indicator variables (or dummy vectors) are computed for each category, resulting in a vector valued $g(z_i)$. The critical ingredient in Eq. (8) is the definition of a suitable *influence function* $h(\cdot)$, which defines the transformation of the responses. To define $h(\cdot)$, we make use of the following proportionality relation of a correlation coefficient $\rho_{jj'}$ with respect to two standardized column vectors $\mathbf{y}^*_{\bullet j}$ and $\mathbf{y}^*_{\bullet j'}$ ($j \neq j'$; $j, j' = 1, \ldots, p$).

$$\rho_{jj'} \propto \langle \mathbf{y}^*_{\bullet j}, \mathbf{y}^*_{\bullet j'} \rangle. \tag{9}$$

The inherent element-wise multiplication of the vector elements $\mathbf{y}^*_{\bullet j} \circ \mathbf{y}^*_{\bullet j'}$ (Hadamard product) results in a new vector $\mathbf{s}_{jj'}$ with n^* elements. Each element gives the contribution of an individual observation to the correlation $\rho_{jj'}$, and is therefore interpretable as score function (cf. Eq. (3) in the MOB approach). For a given i, $h(\mathbf{y}^*_{i\bullet})$ can be expressed as

$$h(\mathbf{y}^*_{i\bullet}) = (y^*_{i1}y^*_{i2}, y^*_{i1}y^*_{i3}, \ldots, y^*_{i2}y^*_{i3}, y^*_{i2}y^*_{i4}, \ldots, y^*_{i(p-1)}y^*_{ip})^\top. \tag{10}$$

Note that here we only include the cross-products for the correlations. If needed, one can also add the identities for the means $y^*_{i1}, y^*_{i2}, \ldots, y^*_{ip}$, and the squared standardized elements for the variances ${y^*_{i1}}^2, {y^*_{i2}}^2, \ldots, {y^*_{ip}}^2$.

The overall matrix structure in Eq. (8) comes from the fact that $h(\mathbf{y}^*_i)$ is always multivariate. Note that the response transformation $g(z_i)$ becomes a matrix only in case of nominal Z. The length of the column vector is therefore not static: it depends on the type of partitioning variable used for the test as well as on the length n of the response vector in a particular node.

To reiterate, $\mathbf{T}$ is used to test for differences of correlations across subgroups ("correlations of correlations"). The distribution of these test statistics has been worked out theoretically on the basis of the permutation principle (Strasser and Weber 1999). The asymptotic permutation distribution of $\mathbf{T}$ is a multivariate normal distribution. This implies that we actually do not have to perform permutations computationally, but rather rely on asymptotic theory (i.e., an infinite amount of permutation samples). $\mathbf{T}$ is computed for each partitioning variable under consideration, resulting in a single p-value for each test. The partitioning variable associated with the lowest p-value is taken as splitting variable, assuming that the p-value is below a pre-specified level α (Bonferroni corrected for the number of partitioning variables).

Now let us assume that a particular continuous Z was selected for splitting. In CTree the split point is chosen such that the corresponding two-sample test statistic in Eq. (8) is maximized. Thus, binary indicator functions $g(z_i)$ coding all possible split points are considered, and then the split is chosen that maximizes the test statistic. Note that in MOB the split is chosen such that the segmented log-likelihood is maximized.

2.3. Some technical and computational remarks

How does MOB differ from CTree within our network context? The overarching difference is that MOB finds a score for a given model (here, a multivariate normal distribution), whereas CTree finds a model for given scores (as defined by the influence function). The difference between the scores $s(\cdot)$ used in MOB and the influence function $h(\cdot)$ in CTree is actually minor. This is due to the fact that ML correlations in a normal model and Pearson correlations are approximately the same (apart from the finite-sample adjustments). Further elaborations on a unifying framework of basic CTree and MOB can be found in Schlosser, Hothorn, and Zeileis (2019).

At a more detailed level, by considering the four steps listed at the beginning of this section, we can say the following. In the MOB approach, all steps are explicitly built on the parametric multivariate normal model, using the corresponding maximum likelihood inference. Thus, the model fit yields the maximum likelihood estimate of all parameters in Step 1. In Step 2 the goodness-of-fit measures are the ML scores which are used for a parameter instability test in Step 3. Step 4 maximizes the partitioned log-likelihood.

In contrast, the approach based on CTree is set up such that the empirical means and covariance matrix are used in Step 1. Note that this is, as mentioned above, equivalent to the ML estimate in a normal model up to potential degrees-of-freedom adjustments. Step 2 then simply uses the corresponding empirical products underlying the correlation matrix as the goodness-of-fit measure. If specified, the means and variances can also be included in this step. In Step 3 an (asymptotic) permutation test based on the relevant parameters is used. The split is selected by maximizing the two-sample permutation test statistic in Step 4.

On a computational note, both MOB and CTree networks are implemented in the networktree package (Jones, Simon, and Zeileis 2020). Through a methods argument the user can choose between MOB or CTree for network partitioning, the former being more efficient in terms of computing time. The networktree package provides the high-level interface for fitting NTs and implements ML estimation and score functions of the multivariate normal distribution. This high-level interface builds on the partykit package (Hothorn and Zeileis 2015) that provides the infrastructure for recursive partitioning using either the MOB or the CTree approach. The MOB/CTree computations in the networktree package use correlations. For plotting the networks in the terminal nodes the user can choose among various types of networks: correlation, partial correlation, or graphical lasso (e.g., Friedman, Hastie, and Tibshirani 2008). Correlation matrices were converted to partial correlation matrices using the following formula: Let $a_{jj'}$ be an element of the inverse variance-covariance matrix $\mathbf{\Sigma}^{-1}$. The partial correlation is given by $\rho_{jj'}^{(\text{partial})} = -a_{jj'}/(\sqrt{a_{jj}}\sqrt{a_{j'j'}})$. Computationally, conversions to partial correlations or graphical lasso networks was implemented via the qgraph package (Epskamp, Cramer, Waldorp, Schmittmann, and Borsboom 2012), which is also used for network visualization in the terminal nodes.

3. Simulations

We are interested in how the power of NT for both MOB and CTree might differ depending on

1. the total sample size n;
2. the number of variables p included in the network;
3. the magnitude of change of a single or multiple values in $\boldsymbol{\rho}$;
4. whether additional spurious partitioning variables $Z_2, \dots, Z_q$ were included.

We are also interested in the type I error when no changes occur, or when changes occur to variable means μ or variable variances $\boldsymbol{\sigma}^2$ but not to correlations $\boldsymbol{\rho}$. Throughout the simulations, we focus on testing for partial structural changes in the correlation parameters $\boldsymbol{\rho}$. To determine estimates for p and n in realistic scenarios, we relied on a review of recent psychometric network analyses by Haslbeck and Fried (2017): values of p ranged from 6 to 69 with a median of 17, and values of n ranged from 63 to 9282 with a median of 420. Based on these values, we chose to simulate at values of $n = 100, 500, 1500$, and $p = 5, 15, 25$.

NTs estimate a $p \times p$ correlation matrix $\mathbf{R}$ containing $p(p-1)/2$ unique parameters $\boldsymbol{\rho}$ for each partition. To estimate $\boldsymbol{\rho}$, the sample size for each partition (n^*) must be at least as large as the number of values in $\boldsymbol{\rho}$ (i.e., $n^* \geq p(p-1)/2$ per partition after splitting). Therefore, the minimum n^* is quadratically larger than p, and can be very large at large values of p. For example, when $p = 5$, a total n of only 20 is potentially sufficient (i.e., $n^* = 10$ for each partition), but when $p = 25$, a total n of 600 or more is required. We only tested situations when n was sufficiently large compared to p to detect at least one split. Thus, plots of simulations shown later are presented in a triangular fashion, with situations with $n^* < p(p-1)/2$ omitted.

We used the *GeneNet* R package (Schaefer, Opgen-Rhein, and Strimmer 2015) to simulate a correlation matrix with a specified n and p. The correlation matrix was created by first generating a random matrix of partial correlations with a specified density of 0.5 (i.e., 50% of edges were nonzero), and then taking the Moore-Penrose generalized inverse of the partial correlation matrix. The full algorithm for matrix generation is outlined in Schaefer and Strimmer (2004). We then simulated two halves of a dataset using this correlation matrix. In the first half, we used the unaltered correlation matrix and specified that all variable means equal zero. In the second half we modified variable correlations ($\Delta\boldsymbol{\rho}$) or variable means ($\Delta\boldsymbol{\mu}$).[2] We then created a partitioning variable Z_1. This variable was drawn from a normal distribution and was ordered such that the median occurred at the split point (i.e., with respect to the row indices of the full data matrix). Subsequently, we ran either MOB or CTree on the simulated dataset to see if a split would be detected based on the Z_1 variable.

[2] Occasionally, altering $\boldsymbol{\rho}$ caused $\mathbf{R}$ to be non-positive-definite, in which case we located an approximate positive definite solution or discarded the case if an approximate case could not be found.

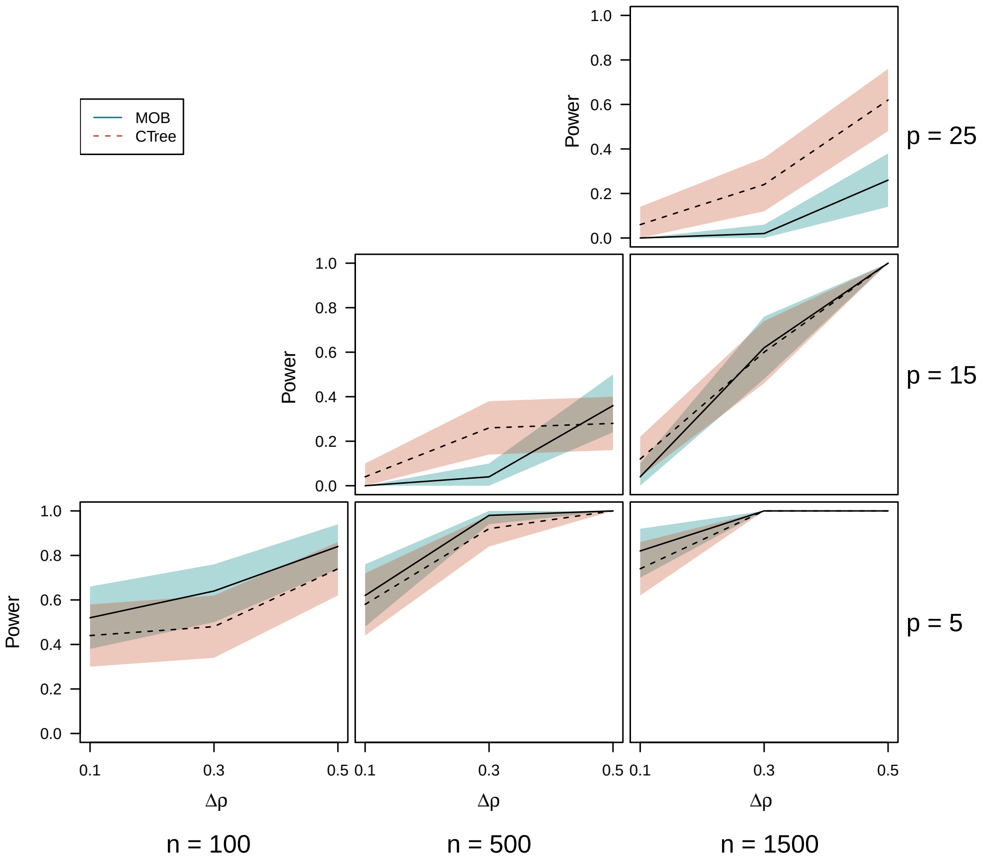

Figure 1: Empirical power for a single change in $\boldsymbol{\rho}$. For each of the six displayed subplots, a single value in $\boldsymbol{\rho}$ was modified as indicated on the x axis. The empirical power is indicated on the y axis with a solid line for MOB and a dashed line for CTree, with shaded 95% confidence intervals. Each subplot corresponds to 50 simulated datasets at specified values of the sample size n and number of nodes p, which vary on the x and y axes of the main plot, respectively.

3.1. Empirical power of NT for a single change in $\boldsymbol{\rho}$

For our first simulation, we altered a single correlation value in $\boldsymbol{\rho}$ by 0.1, 0.3, or 0.5, representing a small, medium, or large correlation according to common effect size conventions. We simulated 50 data sets for each possible combination of n, p, and $\Delta\boldsymbol{\rho}$, which gave us an acceptable trade-off between precision and computing time. We then tested whether MOB/CTree detected a split in the dataset in each case. The results are shown in Figure 1. The x-axis in each plot represents the degree of alteration in a single value in $\boldsymbol{\rho}$. The y-axis represents the empirical power of MOB/CTree at each point, calculated as the proportion of simulated datasets in which a split was found. The performance of MOB is indicated by a solid line and the performance of CTree by a dashed line.

The algorithms were able to consistently detect medium-to-large correlation differences in most scenarios, but performance was lower at $n = 500/p = 15$ and $n = 1500/p = 25$. CTree seemed to have a slight advantage in these scenarios. Unsurprisingly, the algorithms did best at a relatively large n and small p.

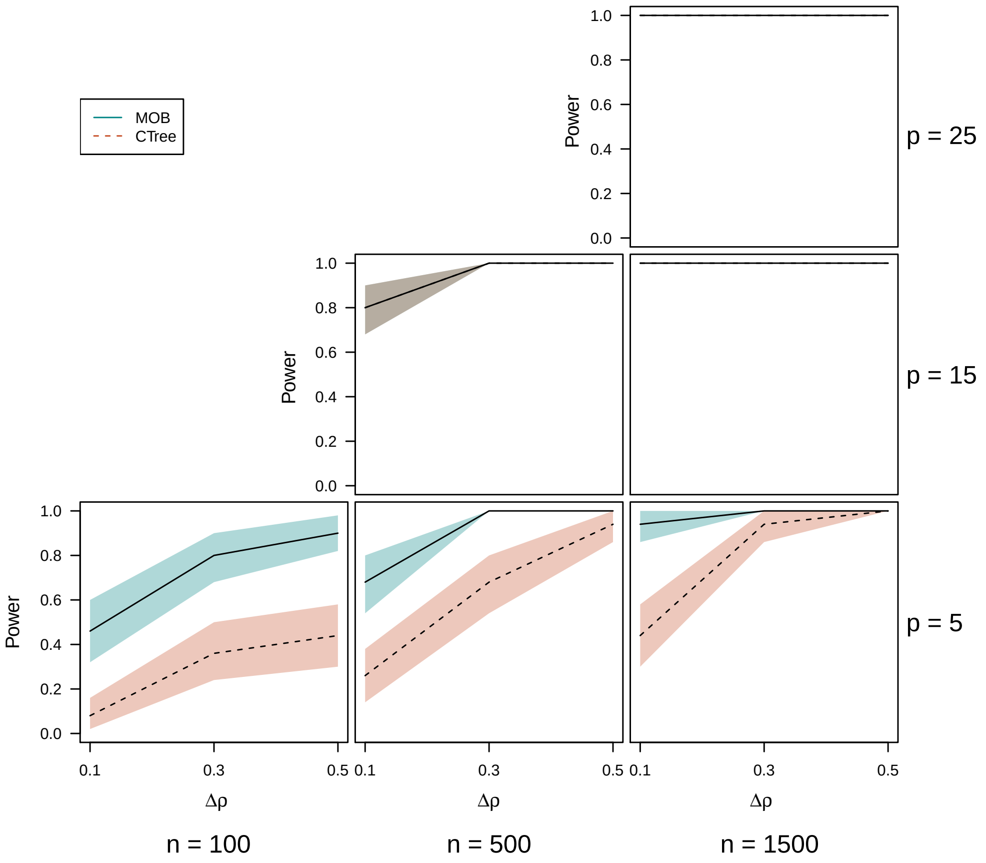

Figure 2: Empirical power when changing all values in $\boldsymbol{\rho}$ For each of the six displayed subplots, all values in $\boldsymbol{\rho}$ were modified as indicated on the x axis. The empirical power is indicated on the y axis with a solid line for MOB and a dashed line for CTree, with shaded 95% confidence intervals. Each subplot corresponds to 50 simulated datasets at specified values of the sample size n and number of nodes p, which vary on the x and y axes of the main plot, respectively.

3.2. Empirical power of NT for multiple changes in $\boldsymbol{\rho}$

We again altered values in $\boldsymbol{\rho}$ by 0.1, 0.3, or 0.5, but instead of altering a single value, we altered all values in $\boldsymbol{\rho}$. The results are shown in Figure 2.

Compared to the first simulation in which a single value in $\boldsymbol{\rho}$ was altered, performance drastically improved in nearly all conditions, especially for large values of p. This indicates that the algorithms are sensitive to the overall sum of change rather than the proportion of change in $\boldsymbol{\rho}$. At small values of p, MOB appeared to show an advantage compared to CTree.

3.3. Empirical power of NT when including spurious partitioning variables

We were interested in whether including spurious partitioning variables that were unrelated to $\boldsymbol{\rho}$ would reduce the power of MOB/CTree. To test this, we repeated the procedure from the first simulation (changing only 1 value in $\boldsymbol{\rho}$) and included 5 spurious partitioning variables $Z_2, \ldots, Z_6$ in addition to the true partitioning variable Z_1.

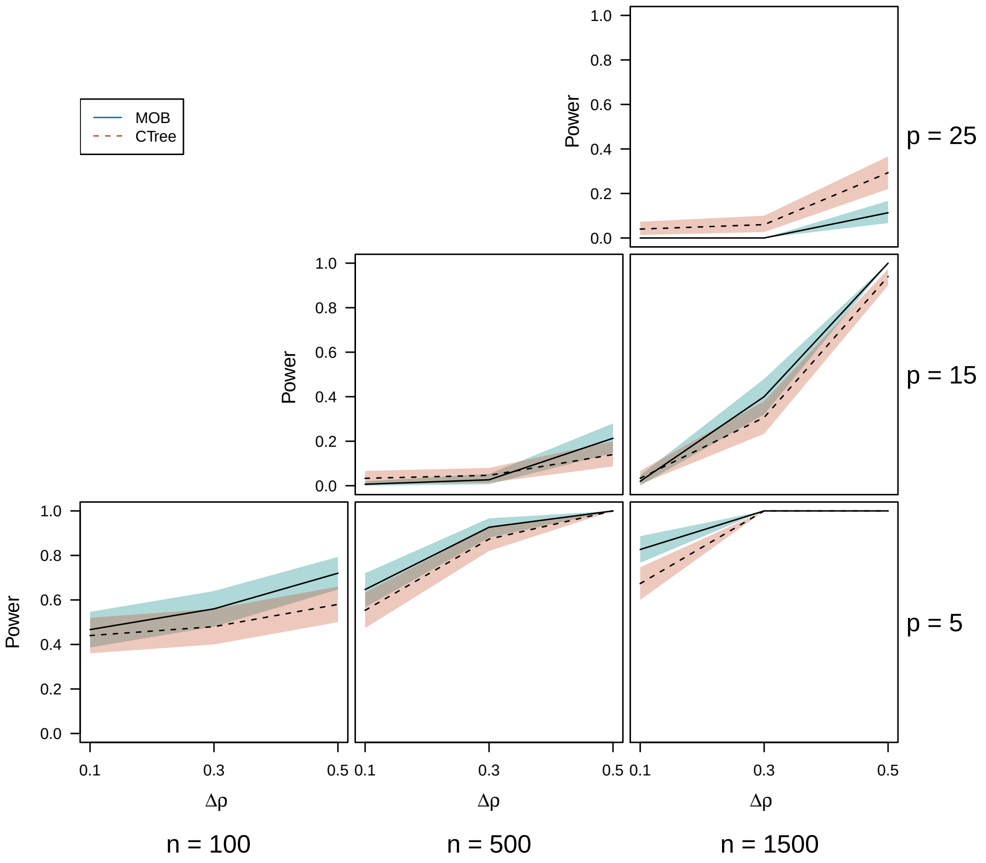

Figure 3: Empirical power when including spurious partitioning variables. For each of the six displayed subplots, a single value in $\boldsymbol{\rho}$ was modified as indicated on the x axis. In addition, five spurious partitioning variables were added to the model. The empirical power is indicated on the y axis with a solid line for MOB and a dashed line for CTree, with shaded 95% confidence intervals. Each subplot corresponds to 150 simulated datasets at specified values of the sample size n and number of nodes p, which vary on the x and y axes of the main plot, respectively.

Results in Figure 3 appeared nearly identical to the first simulation, indicating that including 5 spurious partitioning variables did not meaningfully decrease power.

3.4. False positives: Type I error

Next, we sought to determine whether the algorithms would detect spurious splits when no true splits existed. Thus, we followed the same simulation paradigm as before, but with $\Delta\boldsymbol{\rho} = 0$. We used the same simulated conditions and number of tests per cell as in previous simulations. Here we use k to denote the number of nodes so as not to be confused with p-values.

The results are presented in Table 1. Overall, we found that the type I error rate was acceptably low for MOB and CTree.

Table 1: Type I Error Rate of MOB/CTree. For MOB, results are based on 50 simulated datasets per cell. For CTree, results are based on 300 simulated datasets per cell (increased at the request of a reviewer to improve accuracy).

	MOB			CTree		
	$n = 100$	$n = 500$	$n = 1500$	$n = 100$	$n = 500$	$n = 1500$
$k = 25$			0.00			0.02
$k = 15$		0.02	0.04		0.03	0.04
$k = 5$	0.02	0.02	0.02	0.03	0.04	0.05

Table 2: Detection of a correlation change in MOB (left) and CTree (right) when only the mean was changed. Results are based on 50 simulated datasets per cell.

	MOB			CTree		
	$n = 100$	$n = 500$	$n = 1500$	$n = 100$	$n = 500$	$n = 1500$
$k = 25$			0.00			0.00
$k = 15$		0.02	0.02		0.00	0.04
$k = 5$	0.00	0.02	0.00	0.02	0.00	0.00

Table 3: Detection of a correlation change in MOB (left) and CTree (right) when only the variance was changed. Results are based on 150 simulated datasets per cell.

	MOB			CTree		
	$n = 100$	$n = 500$	$n = 1500$	$n = 100$	$n = 500$	$n = 1500$
$k = 25$			0.00			0.00
$k = 15$		0.00	0.04		0.02	0.05
$k = 5$	0.02	0.00	0.04	0.02	0.03	0.05

3.5. False positives: Change in mean

Relatedly, we were interested if MOB and CTree would detect changes in the correlations when variable means changed, but the correlations remained constant (i.e., $\Delta\boldsymbol{\rho} = 0, \Delta\mu \neq 0$). We repeated the same simulation paradigm as in the above simulations but changed the mean of a single variable by a 10 units (equivalent to 10 standard deviations). The results are presented in Table 2. The results were similar to above ($\Delta\boldsymbol{\mu} = 0$), indicating that changing the mean of a variable does not importantly influence identified splits when testing only for differences in $\Delta\boldsymbol{\rho}$.

3.6. False positives: Change in variance

We were also interested if MOB and CTree would detect changes in the correlations when variable variances changed, but the correlations remained constant (i.e., $\Delta\boldsymbol{\rho} = 0, \Delta\boldsymbol{\sigma}^2 \neq 0$). We repeated the same simulation paradigm as in the above simulations but changed the variance of a single variable by a 10 standard units. The results are presented in Table 3. The results were similar to above ($\Delta\boldsymbol{\sigma}^2 = 0$), indicating that changing the variance of a variable does not importantly influence identified splits when testing only for differences in $\Delta\boldsymbol{\rho}$.

3.7. Simulations: Summary and discussion

These simulations reveal the power and false positive rate of the algorithms under the conditions most common in the psychometric network literature. Unsurprisingly, power typically increased with greater sample size and fewer nodes. The exception to this was when all parameters in the network were altered, in which case more nodes led to a greater chance of detecting at least one split. The power varied greatly depending on the condition, with some conditions (one small effect, 15 nodes, 500 observations) having nearly no power, whereas other conditions (one medium effect, 5 nodes, 1500 observations) had nearly perfect power. The false positive rate for detecting changes in the correlations was acceptable. It remained acceptable when the means or variances were changed.

Our simulations have several limitations. Detecting (or not detecting) a true difference across networks helps us evaluate the power and false positive rate of the splitting parameter, but other parameters are also of interest. Future simulations could also test the precision of the split point in continuous split variables and the accuracy of the covariance parameters of the resultant nodes. Our simulation paradigm used the median as a split point, which represents the best case scenario for sample size proportions across groups; power would be reduced in cases where the sample size is imbalanced. Our simulation approach only evaluated finding a single split in the data; future simulation studies could address situations where multiple hierarchical splits are involved or where other parameters such as network density are manipulated.

Additionally, these simulations only consider whether the algorithm correctly identifies a split. They do not investigate the accuracy of interpretations of subsequent models estimated from the resultant partitions. This concern may be especially important in cases where complex estimation procedures such as regularization are applied to the partitions. For example, we can imagine a scenario where networktree correctly identifies a split, but subsequent regularization shrinks the estimated parameters in each partition to zero. Therefore, if specific parameters estimated from partitions are to be interpreted, post-hoc tests of significance or stability should be provided. As with all post-hoc tests, results should not be interpreted as confirmatory without prespecified hypotheses.

4. Application

In this section, we provide two applications of NT to empirical data using the networktree package. We used data from the Open Source Psychometrics Project[3], an organization that maintains an open website for the public to take psychometric tests for educational and entertainment purposes. All de-identified results are maintained in an open database. For the purposes of this section, we show results from MOB (the package default), indicating any differences with CTree in the text.

For each of the plots in this section, we present partial correlation networks, which are found by taking an inversion of the covariance matrix in each of the terminal nodes. We could also potentially examine the terminal nodes in terms of the raw covariance matrices, correlation matrices, regularized partial correlation networks, factor models, etc. Because the splits are based on the covariance matrix, special caution should be taken when further estimation procedures are applied to the terminal nodes (e.g., graphical lasso networks).

[3] See https://openpsychometrics.org/.

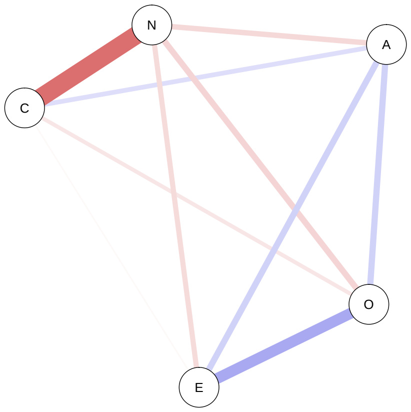

Figure 4: TIPI network from the Open Source Psychometrics Project. Edge thicknesses and transparencies are determined by the strength of partial correlations between nodes. Blue edges represent positive associations and red edges represent negative associations. Node labels correspond to the first letter of each Big Five personality domain (Extraversion, Neuroticism, Conscientiousness, Agreeableness, and Openness to experience).

We selected data from the Ten Item Personality Inventory (TIPI; Gosling, Rentfrow, and Swann Jr 2003), Depression Anxiety and Stress Scale (DASS; Brown, Chorpita, Korotitsch, and Barlow 1997), and accompanying demographic information and validity checks.

First, we examined the TIPI data. The TIPI is a brief inventory of the Big Five personality domains. Scores on each personality domain are calculated by averaging items assigned to each domain (after reverse scoring specific items). After removing individuals that failed validity checks, we were left with 1899 observations. Before applying MOB/CTree, we generated an overall partial correlation network based on the full sample, given in Figure 4. The resulting structure suggests a negative relationship between neuroticism and conscientiousness, a positive relationship between extraversion and openness to experience, and several other potentially interesting relationships.

After gaining a sense of the network structure, we moved into partitioning approaches. We used three potential partitioning variables: “engnat” (Is Engligh your native language? Yes/No), “gender” (What is your gender? Male/Female/Other), and “education” (How much education have you completed? Less than high school/High school/University Degree/Graduate Degree [Ordinal]).

We can draw on the results of our simulations to gain a sense of the power and false positive rate that might be expected from our data structure. As we have 5 nodes and 1899 observations, we can expect to have roughly between 0.75-0.90 power to detect a single small effect with a false positive rate of ≤ 0.05.

The results are shown in Figure 5. The first split occurs between those with a high school degree or less, and those with a university or graduate degree. A second split by education is found, separating those with and without a high school degree. Finally, among those without a high school degree, a split occurs between native and non-native English speakers, perhaps indicating a difference in how the items were being interpreted, or perhaps acting as a proxy for unmeasured cultural variables. Using CTree resulted in the same terminal nodes, although the order of splits differed.

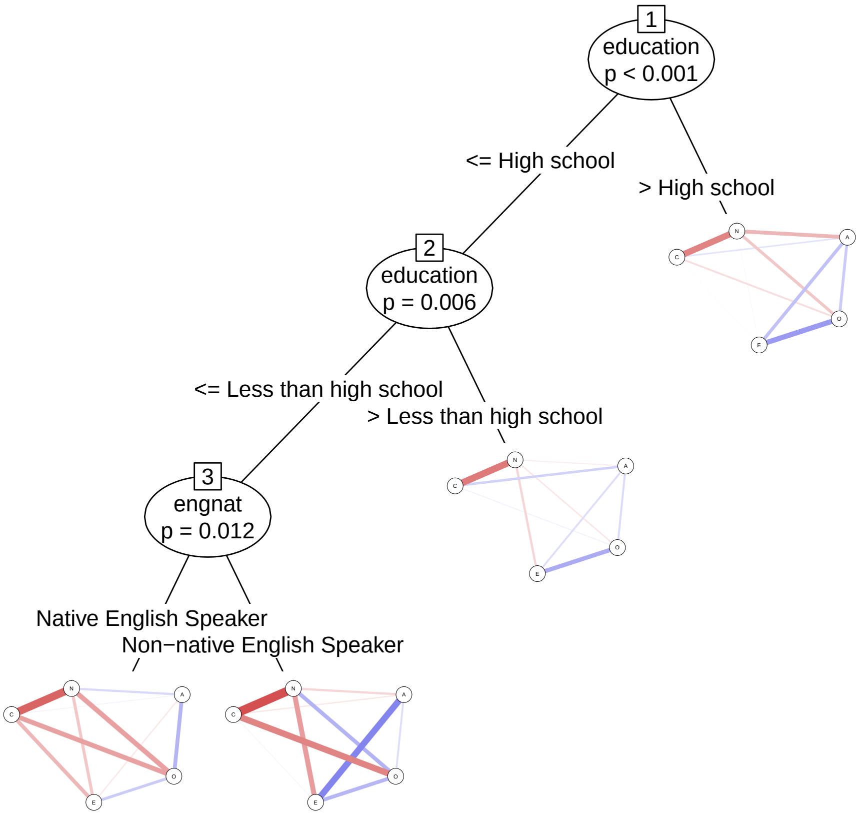

Figure 5: Partitioning TIPI networks. Partitioning was done using MOB. Edge thicknesses and transparencies are determined by the strength of partial correlations between nodes. Node labels correspond to the first letter of each Big Five personality domain (Extraversion, Neuroticism, Conscientiousness, Agreeableness, and Openness to experience).

Importantly, the splits are calculated based on the covariance structure from the data. Depending on one's goals, one may then decide to estimate a specific model using the partitioned data, such as a network model or a factor model. In this example, we converted the covariance matrices of partitions into partial correlation networks. We chose to use partial correlation networks in the examples given their importance in the literature on network theories of psychopathology and personality.

Although the locations of splits are intrinsically informative, the specific parameters in each subnetwork are also likely to be of interest. If these parameters are to be interpreted, appropriate attention should be given to assessing the stability of the subnetwork parameters. This issue is additionally important given that performing network splits diminishes the sample size of each subnetwork. Confidence intervals and p-values can be easily generated for covariances, correlations, and partial correlations. For regularized network models, bootstrapped confidence intervals for edge weights and correlation stability coefficients can be estimated by feeding the subnetwork data into the bootnet package (Epskamp *et al.* 2018). If researchers wish to directly compare parameters across subnetworks, the confidence in these comparisons should also be assessed. Such comparisons can be assessed through a permutation testing framework (Van Borkulo *et al.* 2017) or in a Bayesian context (Williams, Rast, Pericchi, and

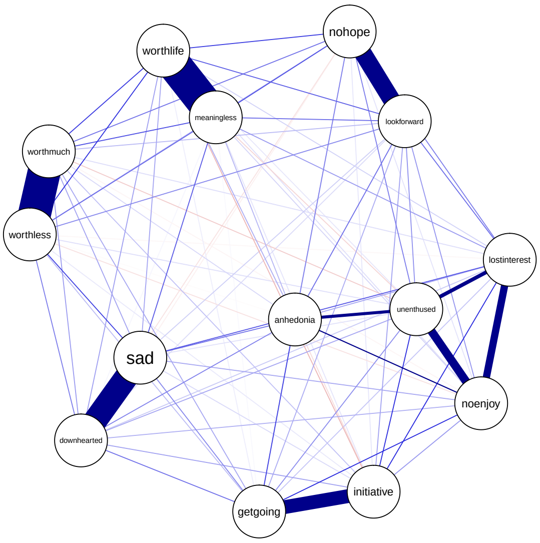

Figure 6: DASS network from the Open Source Psychometrics Project. Edge thicknesses and transparencies are determined by the strength of partial correlations between nodes. Node labels correspond to the depression subscale items of the DASS.

Mulder 2019). Keep in mind that when the network tree approach is used to explore the data without specific a priori hypotheses, any subsequent tests conducted on resultant network structures are exploratory post-hoc tests, and should not be interpreted as confirmatory.

Next, we examined data from the Depression Anxiety and Stress Scale (DASS), a self-report instrument for measuring depression, anxiety, and tension or stress. We focused specifically on the depression subscale of the DASS. We analyzed a random subset of 5000 individuals from the DASS dataset to maintain a more representative scenario. A network of the complete subset is shown in Figure 6. The depression subscale items were: (1) *anhedonia* (I couldn't seem to experience any positive feelings at all), (2) *getgoing* (I just couldn't seem to get going), (3) *lookforward* (I felt that I had nothing to look forward to), (4) *sad* (I felt sad and depressed), (5) *lostinterest* (I felt that I had lost interest in just about everything), (6) *worthmuch* (I felt I wasn't worth much as a person), (7) *worthlife* (I felt that life wasn't worthwhile), (8) *noenjoy* (I couldn't seem to get any enjoyment out of the things I did), (9) *downhearted* (I felt down-hearted and blue), (10) *unenthused* (I was unable to become enthusiastic about anything), (11) *worthless* (I felt I was pretty worthless), (12) *nohope* (I could see nothing in the future to be hopeful about), (13) *meaningless* (I felt that life was meaningless), and (14) *initiative* I found it difficult to work up the initiative to do things.

For the DASS, we used a larger variety of partitioning variables to represent a highly exploratory scenario: "engnat" (Is Engligh your native language? Yes/No), "gender" (What is your gender? Male/Female/Other), "married" (What is your marital status? Never married/Currently married/Previously married), "orientation" (What is your sexual orientation? Heterosexual/Bisexual/Homosexual/Asexual/Other), and "race" (What is your race? Asian/Arab/Black/Indigenous Australian/Native American/White/Other).

The results from the DASS are displayed in Figure 7. The primary split occurred between native and non-native English speakers. Among native English speakers, two further splits were found with the race variable. Among the non-native English speakers, a split was

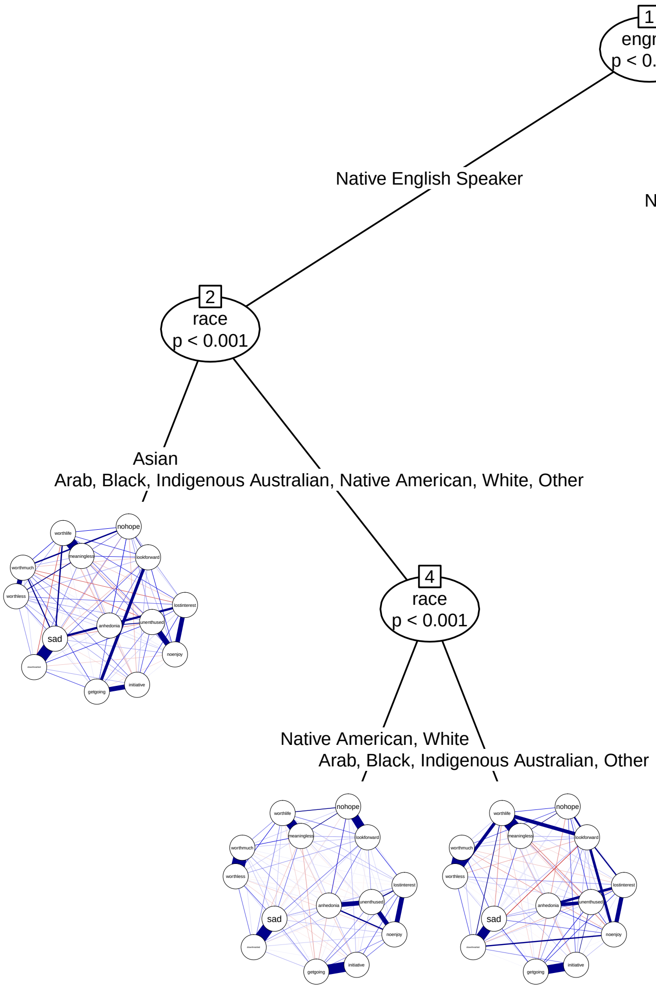

Figure 7: Partitioning DASS networks. Partitioning was done using MOB. Edge thicknesses and transparencies are determined by the strength of partial correlations between nodes. Note: when labels overlap, they are are separated by a line break (left label appears above right label). For long labels, also consider using the print.networktree function to clarify.

found by gender. These results indicate various sources of potential heterogeneity in network structure. Various edge differences differentiated the terminal nodes. For example, among non-native speakers, the connection between *worthlife* (I felt that life wasn't worthwhile) and *nohope* (I could see nothing in the future to be hopeful about) was stronger among males compared to females and other genders. In native English speaking Asians, the connection between *getgoing* (I just couldn't seem to get going) and *lookforward* (I felt that I had nothing to look forward to) was stronger compared to all other racial groups. The results from CTree also showed an initial split between native and non-native English speakers, and several further splits with the race variable among native English speakers, with a split by sexual orientation among non-native English speakers.

As the number of nodes and splits increases, it becomes more difficult to clearly display all relevant parameters in a single graph. Built-in functions in the networktree package facilitate extracting partitioned networks so that they can be easily plotted and examined (see the *getnetwork* function). Other functions facilitate easy comparison of any two given partitions (see the *comparetree* function).

Both applications suggested significant heterogeneity across multiple different partitioning variables. Without explicitly testing for such heterogeneity, it is possible that erroneous conclusions could have been drawn; using network trees elucidates areas where heterogeneity is most impactful. Interestingly, both applications draw attention to the potential role of English as a native language as an important source of potential heterogeneity.

5. Conclusion

The expanding field of network science shows huge promise in psychology through the application of network psychometrics. One of the challenges in the field has been appropriately modeling heterogeneity in network structure using relevant covariates. We propose a recursive partitioning approach to psychometric networks and introduce two techniques that can accomplish this aim, namely MOB and CTree. MOB and CTree recursively split the sample based on covariates in order to detect differences in the covariance matrix. These methods supplement the toolbox of existing techniques such as MNMs and changepoint analyses. Simulations show that both MOB and CTree demonstrate adequate power to detect medium-to-large effects in datasets representative of those most common in the literature.

However, there are some limitations to this method. To identify small effects, large sample sizes are necessary. When considering hierarchical splits, the power is further reduced because the sample size is (roughly) halved at each additional level. For especially large networks, MOB/CTree may not be viable given the quadratic relationship between the number of nodes and the necessary minimum sample size. Future research could address this problem using a regularization approach such as lasso that reduces the number of estimated parameters.

Throughout the paper we treated all variables as metric and applied Pearson correlations. For dichotomous variables, Pearson's ϕ coefficient is a popular measure of association between two variables. Since ϕ is nothing else than the Pearson correlation formula applied on dichotomous variables, the NT methodology works in the same way as above. The scores are still meaningful measures, even though the likelihood is misspecified. Having polytomous items, one can think of applying Spearman's correlation. Since the computation of a Spearman correlation amounts to applying Pearson's formula on the rank version of the data, one can transform the original variable scores into ranks and apply NT as outlined above. Alternatively, other optimal scaling techniques such as correlational aspects (Mair and De Leeuw 2010) can be used to pre-process the data. In order to use tetrachoric, polychoric, or polyserial correlations, further research is needed since these coefficients are typically found in an iterative manner (see, e.g., Drasgow 1986).

In summary, recursive partitioning is a powerful addition to the network science toolkit in psychology and has broad applications for addressing heterogeneity in network structures.

Acknowledgments

P. Jones acknowledges funding from the National Science Foundation (GRFP, grant no. DGE1745303). T. Simon acknowledges funding from the Austrian Science Fund (FWF, grant no. P31836)

References

American Psychiatric Association (2013). *Diagnostic and Statistical Manual of Mental Disorders (DSM-5)*. American Psychiatric Publishing.

Andrews DWK (1993). "Tests for Parameter Instability and Structural Change with Unknown Change Point." *Econometrica*, **61**, 821–856. doi:10.2307/2951764.

Boker SM, Martin M (2018). "A Conversation between Theory, Methods, and Data." *Multivariate Behavioral Research*, **53**(6), 806–819. doi:10.1080/00273171.2018.1437017.

Borsboom D (2017). "A Network Theory of Mental Disorders." *World Psychiatry*, **16**, 5–13. doi:10.1002/wps.20375.

Brandmaier AM, von Oertzen T, McArdle JJ, Lindenberger U (2013). "Structural Equation Model Trees." *Psychological Methods*, **18**(1), 71. doi:10.1037/a0030001.

Breiman L, Friedman JH, Olshen RA, Stone CJ (1984). *Classification and Regression Trees*. Chapman & Hall/CRC, New York.

Brown TA, Chorpita BF, Korotitsch W, Barlow DH (1997). "Psychometric Properties of the Depression Anxiety Stress Scales (DASS) in Clinical Samples." *Behaviour Research and Therapy*, **35**, 79–89. doi:10.1016/s0005-7967(96)00068-x.

Cabrieto J, Tuerlinckx F, Kuppens P, Wilhelm FH, Liedlgruber M, Ceulemans E (2018). "Capturing Correlation Changes by Applying Kernel Change Point Detection on the Running Correlations." *Information Sciences*, **447**, 117–139. doi:10.1016/j.ins.2018.03.010.

Costantini G, Richetin J, Preti E, Casini E, Epskamp S, Perugini M (2019). "Stability and Variability of Personality Networks. A Tutorial on Recent Developments in Network Psychometrics." *Personality and Individual Differences*, **136**, 68–78. doi:10.1016/j.paid.2017.06.011.

Dalege J, Borsboom D, Van Harreveld F, Van den Berg H, Conner M, Van der Maas HL (2016). "Toward a Formalized Account of Attitudes: The Causal Attitude Network (CAN) Model." *Psychological Review*, **123**, 2–22. doi:10.1037/a0039802.

Drasgow F (1986). "Polychoric and Polyserial Correlations." In S Kotz, NL Johnson (eds.), *Encyclopedia of Statistical Sciences*, volume 7, pp. 68–74. John Wiley & Sons, New York.

Epskamp S, Borsboom D, Fried EI (2018). "Estimating Psychological Networks and Their Accuracy: A Tutorial Paper." *Behavior Research Methods*, **50**, 195–212. doi:10.3758/s13428-017-0862-1.

Epskamp S, Cramer AOJ, Waldorp LJ, Schmittmann VD, Borsboom D (2012). "**qgraph**: Network Visualizations of Relationships in Psychometric Data." *Journal of Statistical Software*, **48**(4), 1–18. doi:10.18637/jss.v048.i04.

Epskamp S, Rhemtulla M, Borsboom D (2017). "Generalized Network Psychometrics: Combining Network and Latent Variable Models." *Psychometrika*, **82**, 904–927. doi:10.1007/s11336-017-9557-x.

Fokkema M, Smits N, Zeileis A, Hothorn T, Kelderman H (2018). "Detecting Treatment-Subgroup Interactions in Clustered Data with Generalized Linear Mixed-Effects Model Trees." *Behavior Research Methods*, **50**, 2016–2034. doi:10.3758/s13428-017-0971-x.

Fried EI, Nesse RM (2015). "Depression Is Not a Consistent Syndrome: An Investigation of Unique Symptom Patterns in the STAR*D Study." *Journal of Affective Disorders*, **172**, 96–102. doi:10.1016/j.jad.2014.10.010.

Friedman J, Hastie T, Tibshirani R (2008). "Sparse Inverse Covariance Estimation with the Graphical Lasso." *Biostatistics*, **9**, 432–441. doi:10.1093/biostatistics/kxm045.

Fritz J, Fried EI, Goodyer IM, Wilkinson PO, Van Harmelen AL (2018). "A Network Model of Resilience Factors for Adolescents with and without Exposure to Childhood Adversity." *Scientific Reports*, **8**, 15774. doi:10.31219/osf.io/hvtng.

Gosling SD, Rentfrow PJ, Swann Jr WB (2003). "A Very Brief Measure of the Big-Five Personality Domains." *Journal of Research in Personality*, **37**, 504–528. doi:10.1016/s0092-6566(03)00046-1.

Hanley GP, Iwata BA, McCord BE (2003). "Functional Analysis of Problem Behavior: A Review." *Journal of Applied Behavior Analysis*, **36**, 147–185. doi:10.1901/jaba.2003.36-147.

Hansen BE (1997). "Approximate Asymptotic p Values for Structural-Change Tests." *Journal of Business & Economic Statistics*, **15**, 60–67. doi:10.2307/1392074.

Haslbeck J, Fried EI (2017). "How Predictable Are Symptoms in Psychopathological Networks? A Reanalysis of 18 Published Datasets." *Psychological Medicine*, **47**, 2767–2776. doi:10.1017/s0033291717001258.

Haslbeck JMB, Borsboom D, Waldorp LJ (2019). "Moderated Network Models." *Multivariate Behavioral Research*, **0**, 1–32. doi:10.1080/00273171.2019.1677207.

Hjort NL, Koning A (2002). "Tests for Constancy of Model Parameters over Time." *Nonparametric Statistics*, **14**, 113–132. doi:10.1080/10485250211394.

Hothorn T, Hornik K, Van de Wiel MA, Zeileis A (2006a). "A Lego System for Conditional Inference." *The American Statistician*, **60**, 257–263. doi:10.1198/000313006x118430.

Hothorn T, Hornik K, Zeileis A (2006b). "Unbiased Recursive Partitioning: A Conditional Inference Framework." *Journal of Computational and Graphical Statistics*, **15**, 651–674. doi:10.1198/106186006x133933.

Hothorn T, Zeileis A (2015). "**partykit**: A Modular Toolkit for Recursive Partytioning in R." *Journal of Machine Learning Research*, **16**, 3905–3909.

Jones P, Simon T, Zeileis A (2020). ***networktree**: Recursive Partitioning of Network Models.* R package version 1.0.0, URL https://CRAN.R-project.org/package=networktree.

Jones PJ, Heeren A, McNally RJ (2017). "Commentary: A Network Theory of Mental Disorders." *Frontiers in Psychology*, **8**, 1305. doi:10.3389/fpsyg.2017.01305.

Komboz B, Strobl C, Zeileis A (2018). "Tree-Based Global Model Tests for Polytomous Rasch Models." *Educational and Psychological Measurement*, **78**, 128–166. doi:10.1177/0013164416664394.

Mair P, De Leeuw J (2010). "A General Framework for Multivariate Analysis with Optimal Scaling: The R Package Aspect." *Journal of Statistical Software*, **32**, 1–23. doi:10.18637/jss.v032.i09.

Marsman M, Borsboom D, Kruis J, Epskamp S, Van Bork R, Waldorp LJ, Van der Maas HLJ, Maris G (2018). "An Introduction to Network Psychometrics: Relating Ising Network Models to Item Response Theory Models." *Multivariate Behavioral Research*, **53**, 15–35. doi:10.1080/00273171.2017.1379379.

McNally RJ (2019). "The Network Takeover Reaches Psychopathology." *Behavioral and Brain Sciences*, **42**, e15. doi:10.1017/s0140525x18001073.

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

Merkle EC, Shaffer VA (2011). "Binary Recursive Partitioning Methods with Application to Psychology." *British Journal of Mathematical and Statistical Psychology*, **64**, 161–181. doi:10.1348/000711010x503129.

Merkle EC, Zeileis A (2013). "Tests of Measurement Invariance without Subgroups: a Generalization of Classical Methods." *Psychometrika*, **78**, 59–82. doi:10.1007/s11336-012-9302-4.

Molenaar PCM (2004). "A Manifesto on Psychology as Idiographic Science: Bringing the Person Back into Scientific Psychology, This Time Forever." *Measurement*, **2**, 201–218. doi:10.1207/s15366359mea0204_1.

Park JH, Sohn Y (2019). "Detecting Structural Changes in Longitudinal Network Data." *Bayesian Analysis*, **15**, 133–157. doi:10.1214/19-ba1147.

R Core Team (2020). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Salters-Pedneault K, Tull MT, Roemer L (2004). "The Role of Avoidance of Emotional Material in the Anxiety Disorders." *Applied and Preventive Psychology*, **11**, 95–114. doi:10.1016/j.appsy.2004.09.001.

Schaefer J, Opgen-Rhein R, Strimmer K (2015). ***GeneNet**: Modeling and Inferring Gene Networks*. R package version 1.2.13, URL https://CRAN.R-project.org/package=GeneNet.

Schaefer J, Strimmer K (2004). "An Empirical Bayes Approach to Inferring Large-Scale Gene Association Networks." *Bioinformatics*, **21**, 754–764. doi:10.1093/bioinformatics/bti062.

Schlosser L, Hothorn T, Zeileis A (2019). "The Power of Unbiased Recursive Partitioning: A Unifying View of CTree, MOB, and GUIDE." *arXiv 1906.10179*, arXiv.org E-Print Archive. URL https://arxiv.org/abs/1906.10179.

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

Strasser H, Weber C (1999). "On the Asymptotic Theory of Permutation Tests." *Mathematical Methods of Statistics*, **8**, 220–250.

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

Strobl C, Malley J, Tutz G (2009). "An Introduction to Recursive Partitioning: Rationale, Application, and Characteristics of Classification and Regression Trees, Bagging, and Random Forests." *Psychological Methods*, **14**, 323–348. doi:10.1037/a0016973.

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

Van Borkulo CD, Boschloo L, Kossakowski JJ, Tio P, Schoevers RA, Borsboom D, Waldorp LJ (2017). "Comparing Network Structures on Three Aspects: A Permutation Test." *Journal of Statistical Software*. doi:10.13140/rg.2.2.29455.38569. Forthcoming.

Wang T, Merkle EC, Zeileis A (2014). "Score-Based Tests of Measurement Invariance: Use in Practice." *Frontiers in Psychology*, **5**, 1–11. doi:10.3389/fpsyg.2014.00438.

Wickelmaier F, Zeileis A (2018). "Using Recursive Partitioning to Account for Parameter Heterogeneity in Multinomial Processing Tree Models." *Behavior Research Methods*, **50**, 1217–1233. doi:10.3758/s13428-017-0937-z.

Williams DR, Rast P, Pericchi LR, Mulder J (2019). "Comparing Gaussian Graphical Models with the Posterior Predictive Distribution and Bayesian Model Selection." *PsyArXiv yt386*, PsyArXiv Preprints. doi:10.31234/osf.io/yt386.

Zeileis A (2006). "Implementing a Class of Structural Change Tests: An Econometric Computing Approach." *Computational Statistics & Data Analysis*, **50**, 2987–3008. doi:10.1016/j.csda.2005.07.001.

Zeileis A, Hornik K (2007). "Generalized M-Fluctuation Tests for Parameter Instability." *Statistica Neerlandica*, **61**, 488–508. doi:10.1111/j.1467-9574.2007.00371.x.

Zeileis A, Hothorn T, Hornik K (2008). "Model-Based Recursive Partitioning." *Journal of Computational and Graphical Statistics*, **17**, 492–514. doi:10.1198/106186008x319331.

A. Score functions of the multivariate gaussian distribution

We derive the score functions of the multivariate Gaussion distribution for individual observations. In order to fit the networktree model described in the paper, one only requires the score functions of the correlation parameters. However, for the sake of completeness the score functions for all parameters, i.e., mean, standard deviation, and correlations, are derived. The first derivatives of the log-likelihood, as given in Eq. (2), w.r.t. the parameters μ_k, σ_k, and ρ_{kl} are given below. Note that we use the scalar parameter expressions rather than vector notation. The partial derivative w.r.t. μ_k is:

$$\frac{\partial \ell}{\partial \mu_k} = \sum_{l=1}^{p} \varsigma_{kl}(y_l - \mu_l),$$

where ς_{kl} denotes the element in the k-th row and l-th column of the inverse of the covariance matrix $\mathbf{\Sigma}^{-1}$. The partial derivative w.r.t. σ_k is:

$$\frac{\partial \ell}{\partial \sigma_k} = -\frac{1}{\sigma_k} + \frac{1}{\sigma_k}\left(\frac{y_k - \mu_k}{\sigma_k}\right)\sum_{l=1}^{p} \omega_{kl}\left(\frac{y_l - \mu_l}{\sigma_l}\right),$$

where ω_{kl} denotes the element in the k-th row and l-th column of the inverse of the correlation matrix $\mathbf{R}^{-1}$. Finally, the partial derivative w.r.t. ρ_{kl} is:

$$\frac{\partial \ell}{\partial \rho_{kl}} = -\frac{1}{2}\omega_{ij} + \frac{1}{2}\left(\sum_{m=1}^{p} \omega_{km}\left(\frac{y_m - \mu_m}{\sigma_m}\right)\right)\left(\sum_{m=1}^{p} \omega_{lm}\left(\frac{y_m - \mu_m}{\sigma_m}\right)\right).$$

Note that these are the ones used in our NT approach.

Affiliation:

Payton J. Jones, Patrick Mair
Harvard University
E-mail: payton_jones@g.harvard.edu, mair@fas.harvard.edu
URL: https://scholar.harvard.edu/paytonjones/,
https://scholar.harvard.edu/mair/

Thorsten Simon, Achim Zeileis
Universität Innsbruck
E-mail: Thorsten.Simon@uibk.ac.at, Achim.Zeileis@R-project.org
URL: https://www.uibk.ac.at/acinn/people/thorsten-simon.html,
https://www.zeileis.org/