Article Title

Taxometric Analysis

Author and Co-author Contact Information

John Ruscio (corresponding author)
Psychology Department
The College of New Jersey
2000 Pennington Road
Ewing, NJ 08628
ruscio@tcnj.edu
609-771-2919

Shirley B. Wang
Department of Psychology
Harvard University
33 Kirkland Street, Office 1206
Cambridge, MA 02138
shirleywang@g.harvard.edu
617-495-3800

Abstract

Taxometric analysis is often used in clinical psychological research to determine whether a construct of interest is categorical or dimensional in nature. This chapter reviews the method and provides empirical guidelines for performing and interpreting results of taxometric analysis. Doing so can be quite subjective, and we describe recent advances for reducing this subjectivity. We describe a software package (RTaxometrics) for taxometric analysis and demonstrate its use with illustrative categorical and dimensional data sets. These analyses show how to determine whether data sets are appropriate for taxometric analysis, how to perform various taxometric procedures, and how to interpret the results.

Keywords

taxometric analysis, categories, dimensions, MAMBAC, MAXCOV, MAXEIG, L-Mode, MAXSLOPE, Comparison Curve Fit Index, CCFI, CCFI profile, software, R package

1. Introduction

One major challenge faced by scientists involves determining the latent structure of their variables of interest. Of particular interest to psychologists, people can differ on any given psychological construct by belonging to discrete groups or by varying along a continuum (Meehl, 1992). However, constructs may be conceptualized and measured using either structure based upon theoretical, rather than empirical, grounds. For instance, whereas diagnosis of discrete disorders assumes that individual differences are categorical in nature, evaluating symptom or disorder severity leans towards a dimensional model.

Regardless of any *a priori* preferences, how one chooses to conceptualize and measure a construct and whether this is congruent with its true latent structure has important consequences for theory, research, and practice (Meehl, 1992; Ruscio, Haslam, & Ruscio, 2006; Ruscio & Ruscio, 2002). For instance, knowledge of the structure of a psychological disorder can assist in understanding causal models of psychopathology. Whereas disorders with dimensional variation may be the result of several additive factors (e.g., genetic predisposition, environmental stressors), those that vary in discrete categories require either the presence or absence of a specific causal variable (e.g., a traumatic event) or an accumulation effect, threshold effect, or interaction between variables (e.g., both genetic predisposition and environmental stressors are necessary to push someone over a threshold). Further, structural knowledge can assist researchers in their design and statistical analysis of studies. When measurement models match latent structure, this can increase statistical power of subsequent analyses, such as group comparisons for categorical constructs or tests of association for dimensional constructs (Fraley & Waller, 1998).

Rather than choosing to conceptualize or measure a construct based on conventional practices or preferences for categories or dimensions, this structural distinction can be addressed empirically. Beginning in the 1960s, Paul Meehl and his colleagues published a series of technical reports that introduced a new method for differentiating categorical and dimensional variables. Printed with yellow covers and known informally as the “yellow monsters”, these reports introduced Meehl’s taxometric method. As these reports were circulated and these methods refined, researchers began to incorporate taxometric methodology in their study of psychological constructs. Perhaps because Meehl was a clinical psychologist who developed this method to test for the existence of a schizotype taxon, these methods were largely applied in the realm of psychopathology research. For instance, early taxometric studies examined the latent structure of schizophrenia (Golden & Meehl, 1979), abnormal personality (Erlenmeyer-Kimling, Golden, & Cornblatt, 1989), nuclear depression (Grove et al., 1987), and dementia (Golden, 1982). Reviews of taxometric studies show that they have been used most often to study constructs in the realm of clinical psychology, although researchers have also used these methods to study constructs across all subfields of psychology (e.g., flashbulb memories, infant attachment patterns, emotions) and related fields (e.g., functional dyspepsia, metabolic syndrome) (Haslam, Holland, & Kuppens, 2012; Haslam, McGrath, Viechtbauer, & Kuppens, 2020).

Despite their application to the study of psychopathology and personality, little work was published on the methodology of taxometric analysis until the 1990s. Meehl and Yonce (1994, 1996) illustrated prototypical curve shapes for categorical and dimensional data from analyses of 700 artificial data sets, and this was followed by a demonstration of how to perform several taxometric procedures (Waller, Putnam, & Carlson, 1996). Waller and Meehl (1998) published a book describing existing methods and introducing new procedures. These early methods required a fair amount of subjectivity in that investigators were asked to make a number of choices to implement each taxometric procedure and then visually inspect their taxometric graphs, comparing them to those obtained in analyses of prototypical categorical and dimensional data.

Subsequent developments began to reduce this subjectivity in a variety of ways. Parallel analyses of artificial comparison data provided a clearer sense for what taxometric results would look like for

categorical and dimensional data that reproduced important characteristics of the empirical sample at hand (Ruscio, Ruscio, & Meron, 2007). An objective measure of the relative fit of the obtained results to those for categorical or dimensional data was developed (Ruscio et al., 2007). Simulation studies (e.g., Ruscio, 2007; Ruscio, Carney, Dever, Pliskin, & Wang, 2018; Ruscio & Kaczetow, 2009; Ruscio et al., 2007; Ruscio, Walters, Marcus, & Kaczetow, 2010; Ruscio & Walters, 2011; Walters & Ruscio, 2009) provided further guidance about acceptable data conditions, the best ways to implement taxometric analyses, and the interpretation of results. In the past two decades, the number of taxometric studies has increased rapidly and these methodological safeguards have become standard practice (Haslam et al., 2020). The R package *RTaxometrics* (Ruscio & Wang, 2017) now fully incorporates knowledge on best practices in taxometric analysis in a user-friendly way. This chapter reviews each of these issues in greater detail to help an interested reader assess the merits of a taxometric study or perform one of their own.

2. Overview of the Taxometric Method

At its most fundamental level, taxometric research begins with the premise that not all individual differences are alike. For instance, whereas dogs and cats are qualitatively different in kind, tall people and short people are quantitatively different in degree. However, the latent structure of other constructs is less clear. For instance, do depressed and non-depressed individuals form two separate groups of people, or does everyone fall along a continuous spectrum of depression? Taxometric analysis is designed to address the question of whether a categorical or dimensional model is a better fit for any particular construct. To do so, various taxometric procedures are applied to examine relationships among observable variables for clues to the underlying latent structure.

Within the overarching framework of the taxometric method, dozens of data-analytic procedures have been introduced. A small handful has emerged as the most popular and well-studied set of taxometric procedures. These include mean above minus below a cut (MAMBAC; Meehl & Yonce, 1994), maximum covariance (MAXCOV; Meehl & Yonce, 1996), maximum eigenvalue (MAXEIG; Waller & Meehl, 1998), and latent mode (L-Mode; Waller & Meehl, 1998). The primary output of these procedures is graphical in nature, with certain patterns of graphs more indicative of categorical or dimensional latent structure for each procedure. Although each taxometric procedure is conceptually and mathematically distinct from the others in important ways, they all involve the analysis of multiple valid quantitative indicators of the latent construct (e.g., scores on a depressive symptoms scale as an indicator of depression).

For instance, MAMBAC requires at least two indicator variables (Meehl & Yonce, 1994). First, one indicator is designated as an “input” indicator and the other as an “output” indicator. Scores on the input indicator are used to sort cases. Next, beginning and ending a fixed number of cases away from the lowest and highest scores on the input indicator, a series of cutting scores is located. Mean differences of output indicator scores are calculated above and below each cut. Finally, a MAMBAC graph is created by plotting the series of mean differences corresponding to each cut. A prototypical MAMBAC graph for categorical data shows a peak near the cut that best separates the members of two groups. In contrast, a prototypical MAMBAC graph for dimensional data is concave. Curves for both structures will be shown in the illustrative analyses that appear later. If there are more indicators ($k > 2$), MAMBAC may be repeated $k(k - 1)$ times so that all variables are used as input and output variables to generate a panel of curves, with these curves typically averaged for interpretation.

The MAXCOV and MAXEIG procedures are conceptually very similar to one another, and they yield very similar results (Ruscio et al., 2010). Therefore, we will only describe MAXEIG (Waller & Meehl, 1998). This procedure requires at least three indicator variables. As in MAMBAC, one indicator is designated as an input indicator and cases are sorted along this variable. Ordered subsamples called “windows” of cases are formed such that they overlap, typically by 90%, with their neighbors. Then one calculates, for each window, the first (largest) eigenvalue from a modified variance-covariance

matrix (by replacing the variances with zeros) of all remaining variables, which serve as output indicators. A MAXEIG graph is created by plotting the series of eigenvalues along the mean score of the input indicator for cases in each window. Similar to MAMBAC, categorical data submitted to MAXEIG are expected to yield a peaked curve. Within windows containing mostly members of just one group (e.g., all of the lowest-scoring cases on the input indicator), little association among output indicators is expected. The same holds true within windows containing mostly members of the other group (e.g., all of the highest-scoring cases). When windows contain a fairly even mixture of members of two groups, however, this gives rise to strong associations between indicator variables, hence a peak in the MAXEIG curve. In contrast, a prototypical MAXEIG graph for dimensional data is flat because there are no groups being mixed in differing proportions across windows of cases. Instead, the associations between indicator variables remain fairly constant at all levels of the input indicator.

L-Mode is slightly different than MAMBAC and MAXEIG in that it does not involve cutting or splitting the sample into subgroups (Waller & Meehl, 1998). Instead, all three or more indicators available are submitted to a factor analysis, and scores on the first principal factor are estimated using Bartlett's (1937) method. An L-Mode graph is created by plotting the distribution of cases on this factor as a density plot. Whereas a prototypical L-Mode graph for categorical data is bimodal, revealing the separation between scores for two groups, a prototypical L-Mode graph for dimensional data is unimodal.

Unlike many other forms of latent variable data analyses (e.g., latent class analysis), taxometric procedures do not test for statistical significance to assess the fit of a categorical or dimensional structural model, thereby avoiding the potential pitfalls of null hypothesis statistical testing (Nickerson, 2000; Wagenmakers, 2007). Instead, a cornerstone of Meehl's taxometric method involves checking the consistency of findings across multiple taxometric procedures (Meehl, 1995). These procedures would ideally be applied to multiple datasets drawn from different populations, using different measures as observed indicators of the latent construct. The rationale for consistency testing is not unlike that for replication in other types of research, namely that confidence accumulates only as results from nonredundant tests point toward the same conclusion. Neither a single test nor inconsistent results provide compelling support for an inference of categorical or dimensional latent structure.

3. Reducing Subjectivity in Taxometric Analysis

Research on taxometric methodology has accelerated over the past few decades, with several important advances being made to reduce the subjectivity in taxometric analysis. A key development in this area was the introduction of parallel analyses of artificial comparison data (Ruscio et al., 2007), and this in turn enabled the more rigorous study of taxometric methodology to help decide how best to perform taxometric procedures and interpret their results.

3.1. Parallel Analysis of Comparison Data

Using this approach, one generates populations of categorical and dimensional comparison data by holding constant important characteristics of the empirical data (e.g., sample size, number of variables, marginal distributions, correlation matrices) and varying only the structural models used to create the data. By analyzing many random samples drawn from each population of comparison data, the typical results for each structure can be examined along with the variation attributable to normal sampling error. Plotting results for empirical data alongside those for both types of comparison data provides a more appropriate reference point than comparing the empirical results only to the prototypical curves for each structure that were generated using a narrow range of fairly ideal data parameters.

To further reduce the subjectivity in the interpretation of taxometric results, Ruscio et al. (2007) developed the Comparison Curve Fit Index (CCFI). The CCFI quantifies the extent to which the results

for the empirical data are a closer match to those for the categorical or dimensional comparison data. Values can range from 0 (strongest support for dimensional structure) to 1 (strongest support for categorical structure), with .50 representing the most ambiguous outcome possible. A number of simulation studies demonstrated that the CCFI effectively differentiates between categorical and dimensional data across a wide range of challenging data conditions (see Ruscio, Ruscio, & Carney, 2011, for an overview).

3.2. Inspecting Curves and Curve Fit

Historically, taxometric methodology required investigators to make several judgements about the similarities and differences between graphs for empirical data and prototypical graphs for categorical and dimensional data. These prototypical comparison graphs were generated from a relatively small number of idealized data conditions, which often did not match the distributional and correlational properties of the empirical research data. For instance, whereas empirical data usually differ from normality in one or more ways (Micceri, 1989), the artificial data used to generate the prototypical graphs were normally distributed. Interpreting taxometric results often involved difficult judgments about highly ambiguous comparisons. This reliance on visual inspection of curve shapes introduced an unfortunate degree of subjectivity—and allowed confirmation bias to play an outsized role—in taxometric research.

Compounding this challenge, each taxometric procedure can be performed in a variety ways, and empirical guidance for making implementation decisions was slow to develop because simulation studies required that taxometric experts judge the output of each analysis. Whereas other approaches to latent variable modeling could be studied in large-scale simulation studies using objective measures of model fit, the need to visually inspect curves severely constrained the size and scope of methodological research on the taxometric method.

The use of parallel analyses of categorical and dimensional comparison data, accompanied by the calculation of the CCFI, goes a long way toward addressing these limitations. Graphs generated from artificial comparison data provide a much better interpretative aid by holding constant important characteristics of the data as well as all implementation choices made when performing each taxometric procedure (Ruscio et al., 2007). Calculating the CCFI on the basis of results from these parallel analyses, rather than subjectively interpreting curves shapes relative to idealized prototypes, removes a great deal of subjectivity from taxometric research. Moreover, the CCFI can be used to perform large simulation studies that examine taxometric methodology itself, including questions about necessary data conditions for informative taxometric results as well as the most effective ways to implement taxometric procedures.

Haslam et al. (2012) noted not only that parallel analyses of comparison data and the CCFI have become standard practice in taxometric studies, but also that using the CCFI is strongly associated with higher methodological quality in other respects (e.g., larger sample size, continuous rather than dichotomous indicators). Because the CCFI has become standard practice, Haslam et al. (2020) were able to perform a meta-analysis of taxometric studies using the CCFI as the measure of effect size.

Evidence from Monte Carlo simulation studies shows that the CCFI distinguishes between categorical and dimensional data with a high level of accuracy across various a wide range of challenging data conditions (Ruscio et al., 2007; Ruscio & Kaczetow, 2009; Ruscio et al., 2010; Ruscio et al., 2018). Moreover, using the CCFI allows for the detection of categorical structure with highly unequal base rates of group membership in the sample (Ruscio & Marcus, 2007). This is particularly important as taxometric analysis is frequently applied in the context of psychological disorders, constructs with low base rates.

3.3. Implementation Decisions

Researchers must make a number of implementation decisions when performing a taxometric analysis. For instance, researchers must decide which taxometric procedures to use (e.g., MAMBAC,

MAXEIG, MAXCOV, L-Mode), how to assign variables to input and output configurations, and how to locate cutting scores or subsamples along input variables. In the past, such implementation decisions were made by following conventions suggested in the original papers introducing the methodology or examples in previously published taxometrics studies. Given the many options available to researchers, there was no guarantee that others had made the best choices. The development of the CCFI enabled large-scale simulation studies in which various implementation options were systematically investigated across a wide range of data conditions to uncover acceptable boundary conditions and suggest best practices.

These simulation studies form the foundation of empirically supported guidelines in taxometric analysis. For instance, Ruscio et al. (2010) found that MAXEIG and MAXCOV procedures produced remarkably similar results, and it is now standard practice to only select one of these procedures for use in consistency testing. Other simulation studies have established guidelines for the implementation of MAMBAC, MAXCOV, and MAXEIG procedures (Walters & Ruscio, 2009), as well as the calculation of CCFI values (Ruscio et al., 2018) and use of internal replications when tied scores are found on the input indicator (Ruscio & Walters, 2011).

3.4. CCFI Profiles

A more recent development in taxometric methodology involves performing analyses using a series of populations of categorical comparison data that vary in the base rate of the taxon. The purpose is to examine how the CCFI changes when known groups differ in their relative size. Ruscio et al. (2018) found that creating what they called a CCFI profile using a range of base rates for categorical comparison data (from .025 to .075, in increments of .025) provided two key benefits.

First, using a CCFI profile improves base rate estimation relative to what can be obtained using formulas for each taxometric procedure. If the results support an inference of categorical structure, locating the peak in the CCFI profile provides a clue about the taxon base rate. It is expected that this peak will emerge for the population of categorical comparison data generated using a base rate rate close to that for the empirical data. Because a discrete series of base rates is used to generate the CCFI profile, and also because each CCFI contained therein will be subject to sampling error, the profile is smoothed before locating its peak. The location of the peak in this smoothed curve is then used as the base rate estimate. Ruscio et al. (2018) found that this decreases bias and increases precision of base rate estimation for the MAMBAC, MAXEIG, and L-Mode procedures.

Second, a weighted mean of the CCFI values in a profile improves the ability of CCFI to differentiate between categorical and dimensional data. A single CCFI value is useful, but like any statistic it is subject to sampling error. Averaging values reduces the sampling error and an aggregate CCFI even more effectively differentiates between categorical and dimensional data. The weighting scheme is based on the distance from each data point to the estimate of the taxon base rate, thus giving more weight to points nearer the estimated base rate.

3.5. Consistency Testing

Another cornerstone of Meehl's taxometric method is the use of multiple non-redundant data-analytic procedures to check the consistency of findings (Meehl, 1995). Like other implementation decisions, there are many choices to be made when checking for consistency. The general idea of consistency testing is sound, but with so many "researcher degrees of freedom" (Simmons, Nelson, & Simonsohn, 2011) in selecting which data-analytic techniques to perform and to report, there was a substantial risk of confirmation bias. Indeed, for a long time, researchers' approaches to consistency testing were uneven, at best. Practice was guided only by a shared ideal that had not been operationalized.

Ruscio et al. (2010) used the CCFI to specify and evaluate several operationalizations of consistency testing. The best method among those they tested was to obtain CCFI values using multiple taxometric procedures and then calculate and interpret the mean CCFI. When a single threshold at

.50 is used in this way, there are inevitable errors (i.e., categorical data that yield a CCFI below .50 or dimensional data that yield a CCFI above .50). Findings suggested that the error rate should be low provided that the data are appropriate for taxometric analysis, but users could reduce it further by treating values close to .50 as ambiguous. For instance, treating CCFIs from .40 to .60 as ambiguous, and reaching no conclusion, eliminated most errors. Using a narrower range of ambiguous CCFIs (e.g., from .45 to .55) yielded fewer ambiguous findings, but at the cost of an increase in the error rate. Alternatives to such fixed-width intervals (e.g., intervals based on multiples of the CCFI's standard error) have also been rigorously evaluated via simulation studies, but results indicated that an ambiguous range of CCFI values should be defined using fixed-width intervals (Ruscio et al., 2018). In all of these ways, development of the CCFI and its use in simulation studies have helped to reduce subjectivity and accelerate research in taxometrics by allowing researchers to select the most appropriate analyses, make decisions to perform them most effectively, and report and interpret their results in a more transparent, standardized, and effective fashion.

4. Software for Taxometric Analysis

Mainstream statistical software does not include taxometric analysis, so investigators have created their own special-purpose code through the years. By the time that the use of simulated comparison data became part of standard practice, most investigators seemed to be using Ruscio's (2016) R code, which incorporated that approach. To check our impression that Ruscio's (2016) R code for taxometric analysis had become the most popular, we performed a review of 37 taxometric studies published from 2011 to 2016 using the search term "taxometric analysis" in Google Scholar. In each case, the researchers used Ruscio's taxometric programs. The code was originally written in the commercial S+ language in 2000, and soon thereafter converted for use in the R computing environment.

This code was updated many times, with the results that one might expect of an incremental, evolutionary process. The original formulation and structure remains, buried beneath a variety of add-ons and modifications. The code's growth rendered it increasingly difficult to read or update, much less to reorganize in more modular and efficient ways. Moreover, even as the practice of taxometric analysis began to converge on best practices supported by methodological research, the difficulty of making substantial changes to the inelegant code meant that some outdated options remained and some new techniques had not been incorporated.

Therefore, we completely reworked Ruscio's R code for taxometric analysis to create the R package *RTaxometrics* (Ruscio & Wang, 2017). We followed the modern style conventions of R programming and documentation to produce an R package that is distributed in the standard way, rather than through a personal web site. Though we borrowed parts from the existing code, the *RTaxometrics* package was designed from scratch to have many advantages over the previously distributed code. First, functions in this package were created and tested to be as user-friendly as possible while enabling, encouraging, and in some cases even requiring users to follow best practices. For instance, many aspects of the data can be checked to ensure they are adequate for taxometric analysis prior to running actual taxometric procedures, and additional checks on the fit between the data and the implementation choices are automatically done before any analyses are performed. A newly developed function also allows for the generation and analysis of CCFI profiles.

Second, this package was programmed to be run-time efficient. Perhaps the most significant improvement, from a run-time perspective, involves the generation of comparison data. As noted above, it has become standard practice in taxometric analysis to generate and submit to parallel analysis artificial comparison data (Ruscio et al., 2007). Generating the necessary populations of categorical and dimensional comparison data, from which random samples are taken for parallel analysis, can take as long or longer than performing all of the taxometric analyses. This step used to be done separately for each taxometric procedure, but *RTaxometrics* generates the populations of comparison data only once, storing and using them as needed for multiple taxometric procedures.

Third, *RTaxometrics* provides status updates once a command is run, with progress being reported as various actions are taken. This includes preliminary checks of the data and program parameter specifications, as well as analyses of empirical and comparison data.

Fourth, once analyses are complete, *RTaxometrics* provides streamlined output. The text and graphical output from analyses have been simplified to help users focus on the most important results and incorporate them into their documents. For instance, a single graph sheet is created with the results from all taxometric procedures performed, rather than producing multiple windows with graphs for each procedure separately. These graphs can be displayed on the screen or written directly to either compressed (.jpeg) or high-resolution (.tiff) files. Likewise, the text output can be displayed on screen or diverted directly to a text file.

Fifth, *RTaxomertrics* is much more modular than previous versions, with anything done repeatedly (e.g., calculating CCFIs) handled in its own function and called by higher-order functions as needed. All program parameters are bundled into a single object passed between all functions, making it simple to add or remove elements in future updates. These changes have all improved readability of the code, which is also written and documented in conventional R style. Steps to follow in a taxometric analysis are provided below, followed by several illustrations using *RTaxometrics*.

5. Performing Taxometric Analysis

5.1. Checking the Data

Before performing taxometric analyses, researchers should ensure that this is the right data-analytic tool to address the research question. Taxometric analysis is designed to differentiate between categorical and dimensional data, where dimensional structure consists of one or more latent factors, and categorical structure consists of two separate groups (with potential dimensional variation within one or both). After making this determination, investigators should next check that their data are acceptable for taxometric analysis, which requires that data meet several requirements in order to reach accurate and informative conclusions (Meehl, 1995; Ruscio et al., 2010). These include total sample size ($N \geq 300$), size and base rate of the putative taxon ($n_t \geq 50$ and $P \geq .10$), number of variables ($k \geq 2$), number of ordered categories per variable ($C \geq 4$), between-group validity of each variable ($d \geq 1.25$), and within-group correlations among variables ($r_{wg} \leq .30$). Although it is desirable for data sets to meet each of these requirements, a number of simulation studies have shown that borderline values on some of these criteria, or failure to meet one or more criteria, may be offset by especially favorable characteristics on other criteria in the same data set (Ruscio et al., 2011).

Analyses to check whether data were appropriate for taxometric analysis were previously completed within the functions for taxometric procedures themselves. For example, if one constructed a set of variables and submitted it to taxometric analysis, the output would include information about the between-group validity and within-group correlations of these variables. Incorporating this into the taxometric functions themselves may have been convenient, but it also may have muddied the distinction between checking whether the data are appropriate for analysis and performing the analysis itself. To make this clearer, the *RTaxometrics* package includes a CheckData() function intended to be run before any taxometric procedures. Running CheckData() requires users to assign cases to putative groups, which can be based on prior theory, diagnostic criteria, or a conventionally applied threshold. If no better alternative exists, a base-rate classification may be assigned by running the ClassifyCases() function, which requires only that the base rate of the putative taxon be provided. CheckData() examines and provides output bearing on each of the characteristics listed above. If data do not meet one or more of these requirements, the function provides warning notes in the output (e.g., “This is smaller than the recommended minimum of $N = 300$”).

5.2. Taxometric Procedures

If the data are determined to be acceptable for analysis, researchers should proceed to performing taxometric analysis using the RunTaxometrics() function. Like the CheckData() function, this also requires the provision of a classification variable. The reason for this is that cases must be assigned to groups to generate a population of categorical comparison data. Options for taxometric procedures include MAMBAC, MAXEIG, L-Mode, and MAXSLOPE. The latter procedure, which was not described earlier, is a seldom-used surrogate for MAXCOV or MAXEIG when there are only two indicator variables available for analysis (Grove, 2004; Ruscio & Walters, 2011).

A review of literature on empirically supported guidelines for taxometric analysis was conducted to determine options that the new code should include, as well as appropriate default choices. Although default options exist, most of these can be modified by changing the object containing bundled program parameters.

MAMBAC is automatically run if $k \geq 2$, where k is the number of observed variables submitted to the analysis. Default settings for MAMBAC include variables being used in all input-output pairings (assign.MAMBAC = 1), cuts starting and ending at 25 points from either extreme (n.end = 25), and 50 total cuts (n.cuts = 50). MAXEIG is automatically run if $k \geq 3$, and default settings include each variable serving as an input variable once (assign.MAXEIG = 1) and overlapping windows at .90 (overlap = .90). Because the MAXEIG and MAXCOV procedures produce such similar results (Ruscio et al., 2010) and should not be used as consistency tests, a single function is provided to perform MAXEIG, but not MAXCOV. In the event that only two variables are provided for analysis, MAXSLOPE is performed instead of MAXEIG. L-Mode is automatically run if $k \geq 3$, and default settings include searching for the left mode beyond -.001 (mode.l = -.001) and searching for the right mode beyond .001 (mode.r = .001). Table 1 provides a complete list of options that can be specified, along with default settings and any required minimum or maximum values.

<Table 1 near hear>

If output from RunTaxometrics() indicates that data appear categorical, users may choose to generate a CCFI profile using the RunCCFIProfile() function to estimate the taxon base rate. This function does not require users to provide a classification variable; however, users must still specify procedures and implementation (or rely on default options) as if using RunTaxometrics(). To estimate base rate of the empirical data, RunCCFIProfile() will systematically vary the base rate in the populations of categorical comparison data, displaying CCFI values for each base rate. If this profile is peaked, the location of the peak is used to estimate the base rate for the empirical data (for details, see Ruscio et al., 2018). Of note, this CCFI profile technique can be used either along with or in place of RunTaxometrics(), as it appears to perform as well or slightly better than the conventional approach at differentiating between categorical and dimensional data. However, generating CCFI profiles is considerably more computing- and time-intensive, and it may not be practical to begin with this approach. CCFI profiles are included in the demonstrations to which we turn next.

6. Illustrative Analyses

To demonstrate the use of *RTaxometrics*, we will proceed step-by-step through the analysis of four artificial data sets, including both categorical and dimensional datas that are both unambiguous (idealized data conditions) and ambiguous (some data properties outside the range of conventionally acceptable values). Each of these analyses, including the creation of our illustrative datasets, can be reproduced using *RTaxometrics* and the provided code.

6.1. Unambiguous Categorical Data

6.1.1. Creating the Data

The CreateData() function creates an artificial data set based on either categorical or dimensional structure, including within-group correlations, skew, and/or ordered categorical values if desired

(see Table 2 for full details on default settings and optional parameter specifications for the CreateData() function). This function is useful for becoming familiar with taxometric procedures and the RTaxometrics package, even if one does not have an empirical dataset with which to perform analyses. The program returns a data object containing the variables and a final column containing group membership (1 = complement, 2 = taxon). For dimensional data, this final column is created using the ClassifyCases() function described below, and the codes do not correspond to actual groups. Artificial data can be useful for getting to know the taxometric programs and becoming familiar with their output by conducting analyses using data sets whose characteristics are known.

<Table 2 near here>

First, suppose we wished to create a categorical data set by running the CreateData() function. These data are assigned to the object "x1" so they can be provided to other functions:

```
> x1 <- CreateData("cat", p = .25)
```

By specifying the argument "cat", the function will create a categorical data set. As this function used all default settings (aside from the size of the taxon), this function will create a set of unambiguously categorical data.

6.1.2. Checking the Data

The CheckData() function checks whether the data are appropriate for taxometric analysis. Users should ensure that the data set is a matrix object including one variable per column, followed by a final column containing case classification coded as 1 = complement, 2 = taxon. If the data set does not include this final classification column, users can run the ClassifyCases() function described below to assign cases to groups. Using the first dataset created above, running CheckData() is relatively straightforward:

```
> CheckData(x1)

Sample size:  N = 600
Taxon base rate:  P = 0.25
Taxon size:  n = 150
Complement size:  n = 450
Number of variables:  k = 4

Distributions:

       M   SD Skewness Kurtosis
v1 -0.53 1.35     0.17    -0.24
v2 -0.50 1.38     0.31    -0.26
v3 -0.49 1.31     0.29    -0.15
v4 -0.43 1.36     0.18     0.11

Validities:

     Cohen's d
v1        2.05
v2        1.96
v3        2.01
v4        1.72
Mean      1.93

Within-group correlations (taxon):

      v1    v2    v3    v4
v1  1.00 -0.02 -0.09 -0.03
v2 -0.02  1.00 -0.02 -0.01
v3 -0.09 -0.02  1.00  0.12
v4 -0.03 -0.01  0.12  1.00
```

```
Mean = -0.01

Within-group correlations (complement):

      v1    v2   v3    v4
v1  1.00  0.03 0.03 -0.03
v2  0.03  1.00 0.03 -0.03
v3  0.03  0.03 1.00  0.03
v4 -0.03 -0.03 0.03  1.00
Mean = 0.01
```

If one or more data requirements (e.g., sufficiently large sample size, taxon size, and between-group validity, as well as sufficiently small within-group correlations) are not met, the program will print warnings. In this case, no concerns were noted. Because these data appear adequate for taxometric analysis, we will proceed with the analysis.

6.1.3. Running Taxometric Analyses

The RunTaxometrics() function performs taxometric analyses for a sample of data. If the supplied (empirical) data set contains three or more variables ($k \geq 3$), the function will automatically run the MAMBAC, MAXEIG, and L-Mode procedures. If the supplied data set contains only two variables, the function will automatically run only the MAMBAC and MAXSLOPE procedures. Otherwise, users may also specify which procedures they wish to perform by specifying the MAMBAC, MAXEIG, L-Mode, and MAXSLOPE parameters as TRUE or FALSE. This function requires one argument to be specified, namely the data set. Users may also choose to specify a variety of other shared and procedure-specific parameters (see Table 1 for details). Here, we allow the program to use default settings:

```
> RunTaxometrics(x1)

STATUS OF PROGRAM EXECUTION

Checking for missing data
Checking classification variable
Checking for variance
Checking program parameters
Generating population of dimensional comparison data
Generating population of categorical comparison data
  Generating taxon
  Generating complement
Analyzing empirical data
Analyzing samples of dimensional comparison data
Analyzing samples of categorical comparison data

  Note: Users should run the CheckData() function to evaluate whether
        data appear to be adequate for taxometric analysis.

TAXOMETRIC ANALYSIS RESULTS

Summary of shared analytic specifications
  sample size:  600
  number of variables:  4
  comparison data population size:  1e+05
  comparison data samples:  100
  comparison data taxon base rate:  0.25
  replications:  1

Summary of MAMBAC analytic specifications
  cuts:  50 evenly-spaced cuts beginning 25 cases from either extreme
  indicators:  all possible input-output pairs
  number of curves:  12
```

```
Summary of MAXEIG analytic specifications
  subsamples:  50 windows that overlap 0.9
  indicators:  all possible input-output-output triplets
  number of curves:  12

Summary of L-Mode analytic specifications
  position beyond which to search for left mode:  -0.001
  position beyond which to search for right mode:  0.001

Comparison Curve Fit Index (CCFI)
  MAMBAC:  0.932
  MAXEIG:  0.876
  L-Mode:  0.871
  mean:  0.893

  Note: CCFI values can range from 0 (dimensional) to 1 (categorical).
        The further a CCFI is from .50, the stronger the result.

Base Rate Estimates:
  MAMBAC:  0.311
  MAXEIG:  0.386
  L-Mode:
    based on location of left mode:  0.177
    based on location of right mode:  1
    mean:  0.588
  mean:  0.428

  Note: There is no evidence-based way to use base rate estimates to
        differentiate categorical and dimensional data. They should
        only be used if evidence supports categorical structure.
```

Most of the text output involves status updates as the program executes and notifications of what procedures were performed, and in what ways. Once it has been confirmed that procedures were implemented appropriately, the critical output is the CCFI values and, if the user believes the structure to be categorical, the taxon base rate estimates.

<Figure 1 near here>

The graphical output (see Figure 1) includes panels of curves with results for the empirical data (dark line) superimposed above the results for the categorical comparison data, and then the results for the dimensional comparison data. Results for comparison data sets are summarized by plotting the middle 50% of data points as a gray band and light lines that show the minimum and maximum values. From the graphical output, it appears that the L-Mode procedure missed the clear right mode because the curve was taller at a factor score of 0 (x = 0) than at the right mode (near x = 2). Therefore, before interpreting these results, analyses should be rerun with a program specification of "mode.r = 1" to begin the search for the right mode at x = 1, rather than the default setting of x = .001, which will enable the identification of the right mode near x = 2:

```
> RunTaxometrics(x1, mode.r = 1)

STATUS OF PROGRAM EXECUTION

Checking for missing data
Checking classification variable
Checking for variance
Checking program parameters
Generating population of dimensional comparison data
Generating population of categorical comparison data
  Generating taxon
  Generating complement
Analyzing empirical data
```

```
Analyzing samples of dimensional comparison data
Analyzing samples of categorical comparison data

  Note: Users should run the CheckData() function to evaluate whether
        data appear to be adequate for taxometric analysis.

TAXOMETRIC ANALYSIS RESULTS

Summary of shared analytic specifications
  sample size:  600
  number of variables:  4
  comparison data population size:  1e+05
  comparison data samples:  100
  comparison data taxon base rate:  0.25
  replications:  1

Summary of MAMBAC analytic specifications
  cuts:  50 evenly-spaced cuts beginning 25 cases from either extreme
  indicators:  all possible input-output pairs
  number of curves:  12

Summary of MAXEIG analytic specifications
  subsamples:  50 windows that overlap 0.9
  indicators:  all possible input-output-output triplets
  number of curves:  12

Summary of L-Mode analytic specifications
  position beyond which to search for left mode:  -0.001
  position beyond which to search for right mode:  1

Comparison Curve Fit Index (CCFI)
  MAMBAC:  0.932
  MAXEIG:  0.876
  L-Mode:  0.871
  mean:  0.893

  Note: CCFI values can range from 0 (dimensional) to 1 (categorical).
        The further a CCFI is from .50, the stronger the result.

Base Rate Estimates:
  MAMBAC:  0.311
  MAXEIG:  0.386
  L-Mode:
    based on location of left mode:  0.177
    based on location of right mode:  0.341
    mean:  0.259
  mean:  0.318

  Note: There is no evidence-based way to use base rate estimates to
        differentiate categorical and dimensional data. They should
        only be used if evidence supports categorical structure.
```

<Figure 2 near here>

This new graphical output (see Figure 2) shows that L-Mode now correctly identifies the right mode. In this case, both the text and graphical output support a categorical structure, which is correct: CCFIs are well above .50, the MAMABC and MAXEIG curves contain clear peaks, the L-Mode curve is bimodal, and the curves for empirical data are a much closer match to those for categorical than dimensional comparison data. In addition, adjusting the program settings for L-Mode increased the accuracy of its base rate estimate: .259 is very close the correct value of .25. The mean base rate estimate across procedures, .318, was not as accurate.

6.1.4. Generating a CCFI Profile

Because the results appear categorical, we can generate a CCFI profile in an attempt to improve base rate estimation. To do so, we will run RunCCFIProfile() with the same settings as RunTaxometrics(), save for the exclusion of the classification variable in the 5th and final column of the data matrix:

```
> RunCCFIProfile(x1[,1:4], mode.r = 1)

STATUS OF PROGRAM EXECUTION

Checking for missing data
Checking for variance
Checking program parameters
Analyzing empirical data
Generating population of dimensional comparison data
Analyzing samples of dimensional comparison data
Generating populations of categorical comparison data and analyzing samples
  p = 0.025
  p = 0.05
  p = 0.075
  [base rates from .10 to .95 were removed to conserve space]
  p = 0.975

  Note: Users should run the CheckData() function to evaluate whether
        data appear to be adequate for taxometric analysis.

TAXOMETRIC ANALYSIS RESULTS

Summary of shared analytic specifications
  sample size:  600
  number of variables:  4
  comparison data population size:  1e+05
  comparison data samples:  100
  replications:  1

Summary of MAMBAC analytic specifications
  cuts:  50 evenly-spaced cuts beginning 25 cases from either extreme
  indicators:  all possible input-output pairs
  number of curves:  12

Summary of MAXEIG analytic specifications
  subsamples:  50 windows that overlap 0.9
  indicators:  all possible input-output-output triplets
  number of curves:  12

Summary of L-Mode analytic specifications
  position beyond which to search for left mode:  -0.001
  position beyond which to search for right mode:  1

Aggregate Comparison Curve Fit Index (CCFI)
  mean profile:  0.724
  MAMBAC profile:  0.789
  MAXEIG profile:  0.71
  L-Mode profile:  0.673

  Note: CCFI values can range from 0 (dimensional) to 1 (categorical).
        The further a CCFI is from .50, the stronger the result.
        Aggregate CCFI values are a weighted mean of all CCFI values
        in the profile.

Base Rate Estimates
  mean profile:  0.271
  MAMBAC profile:  0.3
```

```
  MAXEIG profile:  0.243
  L-Mode profile:  0.277

  Note: There is no evidence-based way to use base rate estimates to
        differentiate categorical and dimensional data. They should
        only be used if evidence supports categorical structure.
```

<Figure 3 near here>

The text and graphical output (see Figure 3) are still clearly suggestive of categorical structure. CCFIs are closer to .50 than previous results; this is because constructing a CCFI profile uses fallible classification methods (base-rate classification method; Ruscio, 2009) rather than the perfect classification provided by CreateData() and used in RunTaxometrics(). Indeed, the CCFIs obtained earlier using RunTaxometrics() are unrealistically accurate, as empirical data will not include an infalliable classification variable.

In terms of base rate estimation, RunCCFIProfile() provides a mean profile estimate of .271, which is much closer to the correct value of .25 than was the mean estimate of .318 provided by the RunTaxometrics() procedure. It is worth noting, however, these data are ideal for taxometric analysis. In actual research, empirical data may contain some properties (e.g., sample size, correlation among variables) that are at or below conventionally acceptable thresholds. Therefore, the next demonstration creates and utilizes a set of "messier" data.

6.2. Ambiguous Categorical Data

6.2.1. Creating and Checking the Data

The CreateData() function is used to create a second sample of categorical data, this time specifying parameters to create more challenging data rather than relying on prototypical, idealized values:

```
> x2 <- CreateData("cat", n = 350, k = 4, p = .25, d = 1.5, r.tax = .25,
r.comp = .25, g = .6, h = .15, cuts = 6)
```

The challenges introduced here include a smaller sample size, lower taxon base rate, lower indicator validity, larger within-group correlations, greater asymmetry and tail weight than for normal distributions, and discrete values rather than truly continuous score variation. Next, CheckData() will check the data to determine whether they are appropriate for taxometric analysis:

```
> CheckData(x2)

Sample size:  N = 350
Taxon base rate:  P = 0.2514286
Taxon size:  n = 88
Complement size:  n = 262
Number of variables:  k = 4

Distributions:

      M   SD Skewness Kurtosis
v1 2.38 1.21     0.81     0.16
v2 2.16 1.15     0.93     0.54
v3 2.55 1.35     0.67    -0.40
v4 2.49 1.25     0.71    -0.10

Validities:

     Cohen's d
v1        1.75
v2        1.59
v3        2.06
v4        1.87
```

```
Mean      1.82

Within-group correlations (taxon):

     v1   v2   v3   v4
v1 1.00 0.21 0.18 0.27
v2 0.21 1.00 0.34 0.21
v3 0.18 0.34 1.00 0.23
v4 0.27 0.21 0.23 1.00
Mean = 0.24
  * One or more values above the recommended maximum of r = .30.

Within-group correlations (complement):

     v1   v2   v3   v4
v1 1.00 0.32 0.20 0.23
v2 0.32 1.00 0.33 0.35
v3 0.20 0.33 1.00 0.27
v4 0.23 0.35 0.27 1.00
Mean = 0.28
  * One or more values above the recommended maximum of r = .30.
```

Some warnings are noted in the output of this function to indicate that some of the within-group correlations are large. This documents just one of the challenges noted above, and underscores that these data are more representative of empirical data that investigators submit to taxometric analyses than the unambiguous categorical data examined earlier.

6.2.2. Classifying Cases

To treat this sample as actual research data, the correct classification values provided by CreateData() cannot be used. Rather, we will use the ClassifyCases() function to assign cases to the taxon or complement groups by using a taxon base rate estimate. In this case, we will suppose that this estimate is .30, which represents an imperfect guess based on diagnosis, threshold values, theory, or the like. After assigning cases to groups, we will re-check the data:

```
> x2b <- ClassifyCases(x2[, 1:4], p = .3)

> CheckData(x2b)

Sample size:  N = 350
Taxon base rate:  P = 0.3171429
Taxon size:  n = 111
Complement size:  n = 239
Number of variables:  k = 4

Distributions:

      M   SD Skewness Kurtosis
v1 2.38 1.21     0.81     0.16
v2 2.16 1.15     0.93     0.54
v3 2.55 1.35     0.67    -0.40
v4 2.49 1.25     0.71    -0.10

Validities:

     Cohen's d
v1        2.09
v2        2.29
v3        2.28
v4        2.16
Mean      2.20
```

```
Within-group correlations (taxon):

      v1    v2   v3    v4
v1  1.00 -0.02 0.05  0.14
v2 -0.02  1.00 0.04 -0.01
v3  0.05  0.04 1.00  0.07
v4  0.14 -0.01 0.07  1.00
Mean = 0.05

Within-group correlations (complement):

     v1   v2   v3   v4
v1 1.00 0.12 0.01 0.02
v2 0.12 1.00 0.16 0.17
v3 0.01 0.16 1.00 0.15
v4 0.02 0.17 0.15 1.00
Mean = 0.11
```

Using this classification, the data appear adequate for taxometric analysis.

6.2.3. Running Taxometric Analyses

We will again perform taxometric analysis using RunTaxometrics(), specifying a location for the L-Mode procedure to start searching for the right mode. First, we will run this function using the correct classification of cases to groups (provided by CreateData)):

```
> RunTaxometrics(x2, mode.r = 1)

STATUS OF PROGRAM EXECUTION

Checking for missing data
Checking classification variable
Checking for variance
Checking program parameters
  * tied scores, reps set to 10
  * windows too small, set to N / 10 = 35
Generating population of dimensional comparison data
Generating population of categorical comparison data
  Generating taxon
  Generating complement
Analyzing empirical data
Analyzing samples of dimensional comparison data
Analyzing samples of categorical comparison data

  Note: Users should run the CheckData() function to evaluate whether
        data appear to be adequate for taxometric analysis.

TAXOMETRIC ANALYSIS RESULTS

Summary of shared analytic specifications
  sample size:  350
  number of variables:  4
  comparison data population size:  1e+05
  comparison data samples:  100
  comparison data taxon base rate:  0.251
  replications:  10

Summary of MAMBAC analytic specifications
  cuts:  50 evenly-spaced cuts beginning 25 cases from either extreme
  indicators:  all possible input-output pairs
  number of curves:  12

Summary of MAXEIG analytic specifications
```

```
  subsamples:  35 windows that overlap 0.9
  indicators:  all possible input-output-output triplets
  number of curves:  12

Summary of L-Mode analytic specifications
  position beyond which to search for left mode:  -0.001
  position beyond which to search for right mode:  1

Comparison Curve Fit Index (CCFI)
  MAMBAC:  0.778
  MAXEIG:  0.755
  L-Mode:  0.708
  mean:  0.747

  Note: CCFI values can range from 0 (dimensional) to 1 (categorical).
        The further a CCFI is from .50, the stronger the result.

Base Rate Estimates:
  MAMBAC:  0.388
  MAXEIG:  0.447
  L-Mode:
    based on location of left mode:  0.364
    based on location of right mode:  0.411
    mean:  0.388
  mean:  0.408

  Note: There is no evidence-based way to use base rate estimates to
        differentiate categorical and dimensional data. They should
        only be used if evidence supports categorical structure.
```

<Figure 4 near here>

In this example, both the text and graphical output (see Figure 4) again support a categorical structure. CCFIs are well above .50, and the curves for empirical data more closely match the categorical comparison data. Although these results support a categorical structure, it is noteworthy that the base rate estimates are fairly inaccurate (mean estimate = .408, correct value = .25). As these results were based on an entirely correct classification, which researchers will not have in practice, we will run taxometric analysis again with the fallible classification from ClassifyCases():

```
> RunTaxometrics(x2b, mode.r = 1)

STATUS OF PROGRAM EXECUTION

Checking for missing data
Checking classification variable
Checking for variance
Checking program parameters
  * tied scores, reps set to 10
  * windows too small, set to N / 10 = 35
Generating population of dimensional comparison data
Generating population of categorical comparison data
  Generating taxon
  Generating complement
Analyzing empirical data
Analyzing samples of dimensional comparison data
Analyzing samples of categorical comparison data

  Note: Users should run the CheckData() function to evaluate whether
        data appear to be adequate for taxometric analysis.

TAXOMETRIC ANALYSIS RESULTS
```

```
Summary of shared analytic specifications
  sample size:  350
  number of variables:  4
  comparison data population size:  1e+05
  comparison data samples:  100
  comparison data taxon base rate:  0.317
  replications:  10

Summary of MAMBAC analytic specifications
  cuts:  50 evenly-spaced cuts beginning 25 cases from either extreme
  indicators:  all possible input-output pairs
  number of curves:  12

Summary of MAXEIG analytic specifications
  subsamples:  35 windows that overlap 0.9
  indicators:  all possible input-output-output triplets
  number of curves:  12

Summary of L-Mode analytic specifications
  position beyond which to search for left mode:  -0.001
  position beyond which to search for right mode:  1

Comparison Curve Fit Index (CCFI)
  MAMBAC:  0.835
  MAXEIG:  0.649
  L-Mode:  0.743
  mean:  0.742

  Note: CCFI values can range from 0 (dimensional) to 1 (categorical).
        The further a CCFI is from .50, the stronger the result.

Base Rate Estimates:
  MAMBAC:  0.365
  MAXEIG:  0.462
  L-Mode:
    based on location of left mode:  0.364
    based on location of right mode:  0.411
    mean:  0.388
  mean:  0.405

  Note: There is no evidence-based way to use base rate estimates to
        differentiate categorical and dimensional data. They should
        only be used if evidence supports categorical structure.
```

<Figure 5 near here>

These results appear to support categorical structure just as well as those with the correct classification. Examining the comparison data fit (see Figure 5) and CCFIs, all three procedures support categorical structure, and the mean CCFI is .742. However, the base rate estimation continues to be fairly inaccurate, with a mean estimate of 0.405.

6.2.4. Generating a CCFI Profile

Facing categorical or ambiguous results, researchers might consider generating a CCFI profile. Rather than using a single classification of cases, this technique uses a wide range of taxon base rates to classify cases. Each of these is used to generate a new population of categorical comparison data for parallel analyses, and ultimately a series of CCFI values are calculated. Examining the CCFI profile (a plot of CCFIs by taxon base rates) can provide more accurate base rate estimates if data appear to be categorical, and clearer results if data structure is ambiguous.

```
> RunCCFIProfile(x2[,1:4], mode.r = 1)
```

```
STATUS OF PROGRAM EXECUTION

Checking for missing data
Checking for variance
Checking program parameters
  * tied scores, reps set to 10
  * windows too small, set to N / 10 = 35
Analyzing empirical data
Generating population of dimensional comparison data
Analyzing samples of dimensional comparison data
Generating populations of categorical comparison data and analyzing samples
  p = 0.025
  p = 0.05
  p = 0.075
  [base rates from .10 to .95 were removed to conserve space]
  p = 0.975

  Note: Users should run the CheckData() function to evaluate whether
        data appear to be adequate for taxometric analysis.

TAXOMETRIC ANALYSIS RESULTS

Summary of shared analytic specifications
  sample size:  350
  number of variables:  4
  comparison data population size:  1e+05
  comparison data samples:  100
  replications:  10

Summary of MAMBAC analytic specifications
  cuts:  50 evenly-spaced cuts beginning 25 cases from either extreme
  indicators:  all possible input-output pairs
  number of curves:  12

Summary of MAXEIG analytic specifications
  subsamples:  35 windows that overlap 0.9
  indicators:  all possible input-output-output triplets
  number of curves:  12

Summary of L-Mode analytic specifications
  position beyond which to search for left mode:  -0.001
  position beyond which to search for right mode:  1

Aggregate Comparison Curve Fit Index (CCFI)
  mean profile:  0.635
  MAMBAC profile:  0.718
  MAXEIG profile:  0.541
  L-Mode profile:  0.653

  Note: CCFI values can range from 0 (dimensional) to 1 (categorical).
        The further a CCFI is from .50, the stronger the result.
        Aggregate CCFI values are a weighted mean of all CCFI values
        in the profile.

Base Rate Estimates
  mean profile:  0.335
  MAMBAC profile:  0.351
  MAXEIG profile:  0.27
  L-Mode profile:  0.389

  Note: There is no evidence-based way to use base rate estimates to
        differentiate categorical and dimensional data. They should
        only be used if evidence supports categorical structure.
```

<Figure 6 near here>

The text output and graph (see Figure 6) of this CCFI profile again provide support for a categorical structure, such that the CCFI for the mean profile is still clearly above .50 at .635. However, the base rate estimates have now improved, with a mean of .335 that is closer to the correct value of .25. Therefore, generating a CCFI profile seems to have provided some benefits above and beyond a conventional taxometric analysis for these ambiguous categorical data, particularly in providing a more accurate estimate of the taxon base rate.

6.3. Unambiguous Dimensional Data

6.3.1. Creating and Checking the Data

We will use CreateData() to create a third dataset, this time using all default settings, and check this dataset using CheckData():

```
> x3 <- CreateData("dim")
> CheckData(x3)

Sample size:  N = 600
Taxon base rate:  P = 0.5
Taxon size:  n = 300
Complement size:  n = 300
Number of variables:  k = 4

Distributions:

       M   SD Skewness Kurtosis
v1  0.01 1.00     0.00     0.14
v2 -0.06 1.06    -0.03    -0.16
v3 -0.01 1.04    -0.02     0.00
v4  0.01 1.04     0.06    -0.05

Validities:

     Cohen's d
v1        1.56
v2        1.59
v3        1.70
v4        1.68
Mean      1.63

Within-group correlations (taxon):

     v1   v2   v3   v4
v1 1.00 0.17 0.06 0.13
v2 0.17 1.00 0.07 0.12
v3 0.06 0.07 1.00 0.13
v4 0.13 0.12 0.13 1.00
Mean = 0.11

Within-group correlations (complement):

     v1   v2   v3   v4
v1 1.00 0.21 0.27 0.23
v2 0.21 1.00 0.26 0.21
v3 0.27 0.26 1.00 0.15
v4 0.23 0.21 0.15 1.00
Mean = 0.22
```

Because all the distributional and correlational properties of the data appear adequate for taxometric analysis, we proceed to perform them.

6.3.2. Running Taxometric Analyses

We will run the taxometric analysis using all default settings:

```
> RunTaxometrics(x3)

STATUS OF PROGRAM EXECUTION

Checking for missing data
Checking classification variable
Checking for variance
Checking program parameters
Generating population of dimensional comparison data
Generating population of categorical comparison data
  Generating taxon
  Generating complement
Analyzing empirical data
Analyzing samples of dimensional comparison data
Analyzing samples of categorical comparison data

  Note: Users should run the CheckData() function to evaluate whether
        data appear to be adequate for taxometric analysis.

TAXOMETRIC ANALYSIS RESULTS

Summary of shared analytic specifications
  sample size:  600
  number of variables:  4
  comparison data population size:  1e+05
  comparison data samples:  100
  comparison data taxon base rate:  0.5
  replications:  1

Summary of MAMBAC analytic specifications
  cuts:  50 evenly-spaced cuts beginning 25 cases from either extreme
  indicators:  all possible input-output pairs
  number of curves:  12

Summary of MAXEIG analytic specifications
  subsamples:  50 windows that overlap 0.9
  indicators:  all possible input-output-output triplets
  number of curves:  12

Summary of L-Mode analytic specifications
  position beyond which to search for left mode:  -0.001
  position beyond which to search for right mode:  0.001

Comparison Curve Fit Index (CCFI)
  MAMBAC:  0.391
  MAXEIG:  0.326
  L-Mode:  0.201
  mean:  0.306

  Note: CCFI values can range from 0 (dimensional) to 1 (categorical).
        The further a CCFI is from .50, the stronger the result.

Base Rate Estimates:
  MAMBAC:  0.607
  MAXEIG:  0.581
  L-Mode:
```

```
    based on location of left mode:  0
    based on location of right mode:  0.948
    mean:  0.474
  mean:  0.554

  Note: There is no evidence-based way to use base rate estimates to
        differentiate categorical and dimensional data. They should
        only be used if evidence supports categorical structure.
```

<Figure 7 near here>

These results all clearly suggest dimensional structure. As shown in the graphical output (see Figure 7), the MAMBAC and MAXEIG curves contain no peaks, the L-Mode curve is unimodal, and all curves for empirical data are a much closer match to those for dimensional than categorical comparison data. This is reflected in the CCFIs, which are well below .50 (mean CCFI = .306). Base rate estimates should not be interpreted because these data do not appear to be categorical.

6.3.3. Generating a CCFI Profile

Though it might not be worth the time because the taxometric analysis does not suggest categorical structure and therefore there is no taxon base rate to estimate, we will demonstrate how researchers nonetheless could generate a CCFI profile with these data:

```
> RunCCFIProfile(x3[, 1:4])

STATUS OF PROGRAM EXECUTION

Checking for missing data
Checking for variance
Checking program parameters
Analyzing empirical data
Generating population of dimensional comparison data
Analyzing samples of dimensional comparison data
Generating populations of categorical comparison data and analyzing samples
  p = 0.025
  p = 0.05
  p = 0.075
  [base rates from .10 to .95 were removed to conserve space]
  p = 0.975

  Note: Users should run the CheckData() function to evaluate whether
        data appear to be adequate for taxometric analysis.

TAXOMETRIC ANALYSIS RESULTS

Summary of shared analytic specifications
  sample size:  600
  number of variables:  4
  comparison data population size:  1e+05
  comparison data samples:  100
  replications:  1

Summary of MAMBAC analytic specifications
  cuts:  50 evenly-spaced cuts beginning 25 cases from either extreme
  indicators:  all possible input-output pairs
  number of curves:  12

Summary of MAXEIG analytic specifications
  subsamples:  50 windows that overlap 0.9
  indicators:  all possible input-output-output triplets
  number of curves:  12
```

```
Summary of L-Mode analytic specifications
  position beyond which to search for left mode:  -0.001
  position beyond which to search for right mode:  0.001

Aggregate Comparison Curve Fit Index (CCFI)
  mean profile:  0.378
  MAMBAC profile:  0.394
  MAXEIG profile:  0.385
  L-Mode profile:  0.356

  Note: CCFI values can range from 0 (dimensional) to 1 (categorical).
        The further a CCFI is from .50, the stronger the result.
        Aggregate CCFI values are a weighted mean of all CCFI values
        in the profile.

Base Rate Estimates
  mean profile:  0.975
  MAMBAC profile:  0.975
  MAXEIG profile:  0.975
  L-Mode profile:  0.975

  Note: There is no evidence-based way to use base rate estimates to
        differentiate categorical and dimensional data. They should
        only be used if evidence supports categorical structure.
```

<Figure 8 near here>

As expected, these results (see Figure 8 for CCFI profile graph) also provide clear support for dimensional structure, with a CCFI of .378 for the mean profile.

6.4. Ambiguous Dimensional Data

6.4.1. Creating and Checking the Data

To provide a final demonstration, we will now create dimensional data with suboptimal properties to examine whether taxometric analyses are able to identify dimensional structure under more challenging circumstances. This dataset will be created with substantial positive skew and a modest number of discrete values:

```
> x4 <- CreateData("dim", g = .5, cuts = 6, p = .25)
> CheckData(x4)

Sample size:  N = 600
Taxon base rate:  P = 0.25
Taxon size:  n = 150
Complement size:  n = 450
Number of variables:  k = 4

Distributions:

      M   SD Skewness Kurtosis
v1 2.24 1.14     0.99     0.74
v2 2.29 1.22     1.03     0.76
v3 2.38 1.23     0.91     0.41
v4 2.19 1.15     1.07     0.98

Validities:

     Cohen's d
v1        1.55
v2        1.74
v3        1.90
v4        1.67
```

```
Mean      1.72

Within-group correlations (taxon):

      v1    v2    v3    v4
v1  1.00  0.07 -0.03  0.12
v2  0.07  1.00 -0.03 -0.05
v3 -0.03 -0.03  1.00 -0.15
v4  0.12 -0.05 -0.15  1.00
Mean = -0.01

Within-group correlations (complement):

     v1   v2   v3   v4
v1 1.00 0.22 0.23 0.19
v2 0.22 1.00 0.16 0.17
v3 0.23 0.16 1.00 0.17
v4 0.19 0.17 0.17 1.00
Mean = 0.19
```

Note that specifying a taxon base rate (p = .25) when creating dimensional data will not affect the data themselves, only the classification variable included in the final column of the resulting data object. Although these dimensional data were created to be more challenging, they do appear adequate for taxometric analysis.

6.4.2. Running Taxometric Analyses

Once again, we proceed with a standard taxometric analysis using all default settings:

```
> RunTaxometrics(x4)

STATUS OF PROGRAM EXECUTION

Checking for missing data
Checking classification variable
Checking for variance
Checking program parameters
  * tied scores, reps set to 10
Generating population of dimensional comparison data
Generating population of categorical comparison data
  Generating taxon
  Generating complement
Analyzing empirical data
Analyzing samples of dimensional comparison data
Analyzing samples of categorical comparison data

  Note: Users should run the CheckData() function to evaluate whether
        data appear to be adequate for taxometric analysis.

TAXOMETRIC ANALYSIS RESULTS

Summary of shared analytic specifications
  sample size:  600
  number of variables:  4
  comparison data population size:  1e+05
  comparison data samples:  100
  comparison data taxon base rate:  0.25
  replications:  10

Summary of MAMBAC analytic specifications
  cuts:  50 evenly-spaced cuts beginning 25 cases from either extreme
  indicators:  all possible input-output pairs
  number of curves:  12
```

```
Summary of MAXEIG analytic specifications
  subsamples:  50 windows that overlap 0.9
  indicators:  all possible input-output-output triplets
  number of curves:  12

Summary of L-Mode analytic specifications
  position beyond which to search for left mode:  -0.001
  position beyond which to search for right mode:  0.001

Comparison Curve Fit Index (CCFI)
  MAMBAC:  0.409
  MAXEIG:  0.273
  L-Mode:  0.31
  mean:  0.33

  Note: CCFI values can range from 0 (dimensional) to 1 (categorical).
        The further a CCFI is from .50, the stronger the result.

Base Rate Estimates:
  MAMBAC:  0.412
  MAXEIG:  0.149
  L-Mode:
    based on location of left mode:  0.255
    based on location of right mode:  1
    mean:  0.627
  mean:  0.396

  Note: There is no evidence-based way to use base rate estimates to
        differentiate categorical and dimensional data. They should
        only be used if evidence supports categorical structure.
```

<Figure 9 near here>

As shown in the graphical output (see Figure 9), the results for empirical data appear strange. None of the curve shapes approximate prototypes well. The MAMBAC curve appears wavy rather than peaked or concave, the MAXEIG curve is knotty rather than peaked or flat, and the L-Mode curve is generally unimodal, but a bit lumpy. Relying on visual inspection of these curves might yield ambiguous or inaccurate conclusions about which reasonable people could disagree. In this way, we see that the curves for comparison data help to clarify that the empirical data results are a better fit for the dimensional data. Likewise, the CCFI values provide helpful information, with a mean CCFI of .330 indicating stronger support for dimensional data. This underscores the usefulness of comparison data and the CCFI in taxometric analysis (Ruscio & Marcus, 2007; Ruscio & Walters, 2009; Ruscio et al., 2010, 2018).

7. Concluding Remarks

Originally developed by Meehl in the 1960s to test his model of schizotaxia and the development of schizophrenia, research on the methodology and applications of taxometric analysis has rapidly progressed over the past few decades. A major innovation in taxometric methodology was the introduction of comparison data and the CCFI by Ruscio et al. (2007), which prompted a series of Monte Carlo studies that yielded important information about best practices in implementing taxometric procedures. More recently, Ruscio et al. (2018) introduced the CCFI profile, a novel technique that rigorously tests for the existence of groups in empirical data and estimates their size with less bias and greater precision than conventional techniques.

Although taxometric analysis has been most widely applied in the realm of clinical psychology and psychopathology, we also see great potential for this analysis in other fields. For instance, some researchers have begun to apply these methods in social psychology to examine emotions and

emotional/affective processes (e.g., Falcon, 2015), as well as in cognitive psychology to examine the latent structure of secure base script knowledge (Waters et al., 2015) and flashbulb memories (Lanciano & Curci, 2012). In the field of neuroscience, Tran, Stieger, and Voracek (2014) used taxometric analysis to study the latent structure of cerebral lateralization. Future research in these and other areas could provide important information about whether individual differences on any construct of interest are better conceptualized as categories or dimensions at the latent level. As research using taxometric analysis continues to proliferate, we hope that researchers in the psychological, behavioral, and brain sciences will consider whether taxometric analysis could be used to answer meaningful questions in their programs of research.

We encourage researchers who seek to conduct taxometric analysis, or are simply interested in familiarizing themselves with this methodology, to explore the *RTaxometrics* package (available at https://cran.r-project.org/web/packages/RTaxometrics/index.html). This replaces Ruscio's (2016) code, which has been retired, and incorporates all of the methodological advances described in this chapter. *RTaxometrics* is more easily modified and updated, more modular, more readable, and more efficient in the execution of functions and procedures, as well as providing more user-friendly output. We hope that the overview of taxometric methodology and demonstrations in this chapter enable readers to think critically about taxometric studies they encounter in their research and, if interested, to perform their own taxometric analyses.

References

Bartlett, M. S. (1937). The statistical conception of mental factors. *British Journal of Psychology*, *28*(1)*,* 97-104.

Erlenmeyer-Kimling, L., Golden, R. R., & Cornblatt, B. A. (1989). A taxometric analysis of cognitive and neuromotor variables in children at risk for schizophrenia. *Journal of Abnormal Psychology*, *98*(3), 203-208.

Falcon, R. G. (2015). Is envy categorical or dimensional? An empirical investigation using taxometric analysis. *Emotion*, *15*(6), 694-698.

Fraley, R. C., & Waller, N. G. (1998). Adult attachment patterns: A test of the typological model. In J. A. Simpson & W. S. Rholes (Eds.), *Attachment theory and close relationships* (pp. 77-114). New York: Guilford.

Golden, R. R. (1982). A taxometric model for the detection of a conjectured latent taxon. *Multivariate Behavioral Research*, *17*(3), 389-416.

Golden, R. R., & Meehl, P. E. (1979). Detection of the schizoid taxon with MMPI indicators. *Journal of Abnormal Psychology*, *88*(3), 217-233.

Grove, W. M. (2004). The MAXSLOPE taxometric procedure: Mathematical derivation, parameter estimation, consistency tests. *Psychological Reports*, *95*(2), 517-550.

Grove, W. M., Andreasen, N. C., Young, M., Endicott, J., Keller, M. B., Hirschfeld, R. M. A., & Reich, T. (1987). Isolation and characterization of a nuclear depressive syndrome. *Psychological Medicine*, *17*(2), 471-484.

Haslam, N., Holland, E., & Kuppens, P. (2012). Categories versus dimensions in personality and psychopathology: a quantitative review of taxometric research. *Psychological Medicine*, *42*(5), 903-920.

Haslam, N., McGrath, M. J., Viechtbauer, W., & Kuppens, P. (2020). Dimensions over categories: a meta-analysis of taxometric research. *Psychological Medicine*, 1-15.

Lanciano, T., & Curci, A. (2012). Type or dimension? A taxometric investigation of flashbulb memories. *Memory*, *20*(2), 177-188.

Marcus, D. K., Sawaqdeh, A., & Kwon, P. (2014). The latent structure of generalized anxiety disorder in midlife adults. *Psychiatry Research, 215*(2), 366-371.

Meehl, P. E. (1992). Factors and taxa, traits and types, differences of degree and differences in kind. *Journal of Personality, 60*(1)*,* 117-174.

Meehl, P. E. (1995). Bootstraps taxometrics: Solving the classification problem in psychopathology. *American Psychologist*, *50*(4), 266-275.

Meehl, P. E., & Yonce, L. J. (1994). Taxometric analysis: I. Detecting taxonicity with two quantitative indicators using means above and below a sliding cut (MAMBAC procedure). *Psychological Reports*, *74*(3, Pt 2), 1059-1274.

Meehl, P. E., & Yonce, L. J. (1996). Taxometric analysis: II. Detecting taxonicity using covariance of two quantitative indicators in successive intervals of a third indicator (Maxcov procedure). *Psychological Reports*, *78*(3, Pt 2), 1091-1227.

Nickerson, R. S. (2000). Null hypothesis significance testing: a review of an old and continuing controversy. *Psychological Methods*, *5*(2), 241-301.

Ruscio, J. (2007). Taxometric analysis: An empirically-grounded approach to implementing the method. *Criminal Justice and Behavior, 34*(12)*,* 1588-1622.

Ruscio, J. (2009). Assigning cases to groups using taxometric results: An empirical comparison of classification techniques. *Assessment*, *16*(1), 55-70.

Ruscio, J. (2016). Taxometric programs for the R computing environment: User's manual. Available on the world wide web at http://ruscio.pages.tcnj.edu/quantitative-methods-program-code/

Ruscio, J., Carney, L., Dever, L., Pliskin, M., Wang, S.B. (2018). Using the Comparison Curve Fit Index (CCFI) in taxometric analyses: Averaging curves, standard errors, and CCFI profiles. *Psychological Assessment, 30*(6), 744-754.

Ruscio, J., Haslam, N., & Ruscio, A. M. (2006). *Introduction to the taxometric method: A practical guide*. Mahwah, NJ: Lawrence Erlbaum Associates.

Ruscio, J., & Kaczetow, W. (2009). Differentiating categories and dimensions: Evaluating the robustness of taxometric analyses. *Multivariate Behavioral Research*, *44*(2), 259-280.

Ruscio, J., & Marcus, D. K. (2007). Detecting small taxa using simulated comparison data: A reanalysis of Beach, Amir, and Bau's (2005) data. *Psychological Assessment*, *19*(2), 241-246.

Ruscio, J., & Ruscio, A. M. (2002). A structure-based approach to psychological assessment: Matching measurement models to latent structure. *Assessment, 9*(1), 4-16.

Ruscio, J., Ruscio, A. M., & Carney, L. M. (2011). Performing taxometric analysis to distinguish categorical and dimensional variables. *Journal of Experimental Psychopathology*, *2*(2), 170-196.

Ruscio, J., Ruscio, A. M., & Meron, M. (2007). Applying the bootstrap to taxometric analysis: Generating empirical sampling distributions to help interpret results. *Multivariate Behavioral Research*, *42*(2), 349-386.

Ruscio, J., & Walters, G. D. (2011). Differentiating categorical and dimensional data with taxometric analysis: are two variables better than none? *Psychological Assessment*, *23*(2), 287-299.

Ruscio, J., Walters, G. D., Marcus, D. K., & Kaczetow, W. (2010). Comparing the relative fit of categorical and dimensional latent variable models using consistency tests. *Psychological Assessment*, *22*(1), 5-21.

Ruscio, R., Wang, S.B. (2017). RTaxometrics: Taxometric Analysis. R package version 2.3. https://CRAN.R-project.org/package=RTaxometrics

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

Tran, U. S., Stieger, S., & Voracek, M. (2014). Evidence for general right-, mixed-, and left-sidedness in self-reported handedness, footedness, eyedness, and earedness, and a primacy of footedness in a large-sample latent variable analysis. *Neuropsychologia*, *62*, 220-232.

Wagenmakers, E. J. (2007). A practical solution to the pervasive problems of p values. *Psychonomic Bulletin and Review*, *14*(5), 779-804.

Waller, N. G., & Meehl, P. E. (1998). *Multivariate taxometric procedures: Distinguishing types from continua*. Thousand Oaks, CA, US: Sage Publications, Inc.

Waller, N., Putnam, F. W., & Carlson, E. B. (1996). Types of dissociation and dissociative types: A taxometric analysis of dissociative experiences. *Psychological Methods*, *1*(3), 300-321.

Walters, G. D., & Ruscio, J. (2009). To sum or not to sum: Taxometric analysis with ordered categorical assessment items. *Psychological Assessment*, *21*(1), 99-111.

Waters, T. E., Fraley, R. C., Groh, A. M., Steele, R. D., Vaughn, B. E., Bost, K. K., ... & Roisman, G. I. (2015). The latent structure of secure base script knowledge. *Developmental Psychology*, *51*(6), 823-830.

Figures and Tables

Tables are provided in separate Word files.

Figures are provided in separate .tiff files with 500 dpi resolution.

Figure captions are listed below.

Figure 1. Graphs for unambiguous categorical data. Curves for the empirical data are very clearly a closer match for the categorical than dimensional comparison data curves. However, the L-Mode procedure missed the clear right mode because the curve was taller at *x* = 0 than at the ride mode (near *x* = 2).

Figure 2. Graphs for unambiguous categorical data, with data analytic parameters adjusted for L-Mode to begin searching for the right mode at *x* = 1. This adjustment allowed L-Mode to correctly identify the second (right) mode.

Figure 3. CCFI profile for unambiguous categorical data. M = MAMBAC, X = MAXEIG, L – L-Mode, and circles are mean values across these procedures. These curves are clearly suggestive of categorical data, as the CCFIs are consistently above .5, and the peak of the mean profile suggests a base rate of .271.

Figure 4. Graphs for ambiguous categorical data with correct classification of empirical data (based on CreateData()). Curves for the empirical data are very clearly a closer match for the categorical than dimensional comparison data curves.

Figure 5. Graphs for ambiguous categorical data with fallible classification of empirical data (based on ClassifyCases()).Empirical data curves for all three procedures appear closer to categorical comparison data than dimensional comparison data. These results appear similarly ambiguous as those with correct classification (see Figure 4).

Figure 6. CCFI profile for ambiguous categorical data. Results suggest that the data are categorical, as CCFIs appear to be consistently above the .5 threshold.

Figure 7. Graphs for unambiguous dimensional data. MAMBAC and MAXEIG curves contain no peaks, and the L-Mode curve is unimodal. Curves for the empirical data are very clearly a closer match for the dimensional than categorical comparison data curves.

Figure 8. CCFI profile for unambiguous dimensional data. These curves are clearly suggestive of dimensional data, as the CCFIs are consistently below .5.

Figure 9. Graphs for ambiguous dimensional data. The curves for empirical data appear somewhat strange, and do not approximate prototypes well. The MAMBAC curve appears wavy (rather than peaked or concave), the MAXEIG curve appears knotty (rather than peaked or flat), and the L-Mode curve appears generally unimodal, but a bit lumpy. The curves for comparison data help to clarify results, as the empirical data are a better fit for the dimensional than categorical comparison data.

Permissions

N/A

Table 1. Taxometric Program Parameters

Parameter	Function and Default Value
seed	The random number seed provided prior to analysis of empirical data as well as prior to generating each population of comparison data (if comparison data are used); this allows users to create exact replications of analyses. The default value is 1.
n.pop	The size of populations of comparison data. The default value is 100,000, and the minimum value is 10,000.
n.samples	The number of samples drawn from each population of comparison data; Generating multiple sets of comparison data is strongly encouraged. The default value is 100, and the minimum value is 10.
reps	The number of times to resort tied scores and redo calculations, which are averaged to obtain final results. If no tied scores are found, the default and minimum values are 1; if tied scores are found, the default and minimum values are 10.
min.p	The minimum base rate used for generating a CCFI profile. The default value is .025, and the minimum value is 0.025.
max.p	The maximum base rate used for generating a CCFI profile. The default value is .975, and the maximum value is .975.
num.p	The number of base rates used for generating a CCFI profile. The default value is 39, and the minimum value is 20.
MAMBAC	Whether the MAMBAC procedure is performed (default = TRUE).
assign.MAMBAC	Whether the variables are used in all input-output pairings (assign.MAMBAC = 1) or one variable at a time is used as the output variable with all remaining variables summed to form the corresponding input variable (assign.MAMBAC = 2). The default value is 1.
n.cuts	The number of cuts along the input variable in a MAMBAC analysis. The default value is 50, and the minimum value is 25.
n.end	The number of cases at each extreme along the input variable before making the first and last cuts in a MAMBAC analysis. The default value is 25, and the minimum value is 10.
MAXEIG	Whether the MAXEIG procedure is performed (default = TRUE).
assign.MAXEIG	Whether the variables are used in all input-output triplets (assign.MAXEIG = 1), each variable serves as input once with all remaining variables serving as the correspond output variables (assign.MAXEIG = 2), or two variables at a time are used as the output variables with all remaining variables summed to form the corresponding input variable (assign.MAXEIG = 3). The default value is 1.
windows	The number of overlapping windows in a MAXEIG analysis. The default value is 50, and the minimum value is 10.

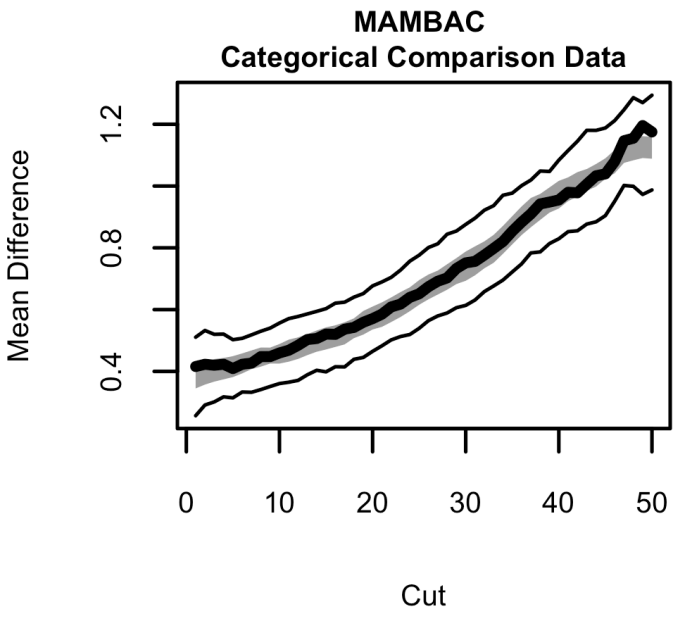
overlap	The proportion of overlap between windows in a MAXEIG analysis. The default value is .90, and the minimum value is 0.
LMode	Whether the L-Mode procedure is performed (default = TRUE).
mode.l	The position beyond which to search for the left mode in an L-Mode analysis. The default value is -.001, and this value must be a negative number.
mode.r	The position beyond which to search for the right mode in an L-Mode analysis. The default value is .001, and this value must be a positive number.
MAXSLOPE	Whether the MAXSLOPE procedure is performed (default = FALSE).
graph	Whether to display the graphical output on screen (graph = 1), to save a compressed .jpeg file (500 dpi, 50% quality; graph = 2), or to save an uncompressed .tiff file (500 dpi; graph = 3).

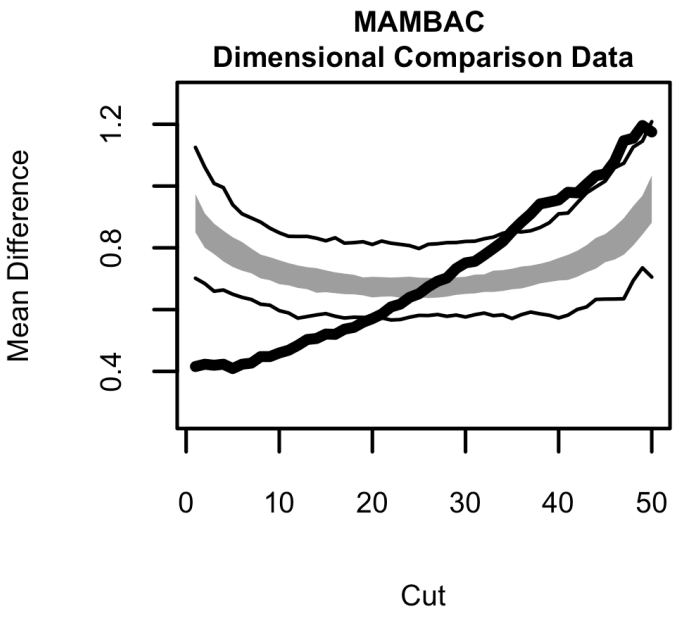
Notes. The parameters for the RunTaxometrics() and RunCCFIProfile() functions are shared across all taxometric procedures (MAMBAC, MAXEIG, L-Mode, MAXSLOPE). All subsidiary functions will automatically run with the defaults shown here, unless otherwise specified by users. Although there is flexibility in adjusting these parameters, some minimum and maximum values are often required. For example, the minimum size of populations of comparison data is 10,000; if users set n.pop to a value less than 10,000, it will automatically be reset to 10,000 (and the user will be notified of this change).

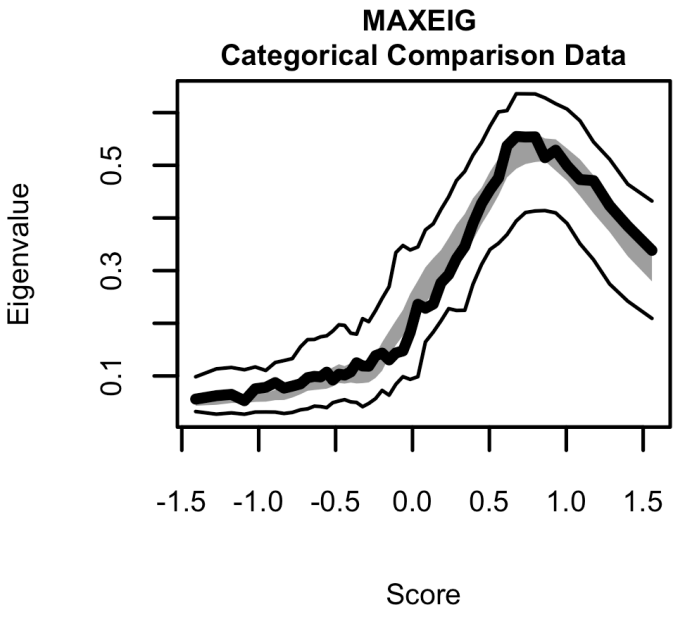
Table 2. Parameters for Creating Data

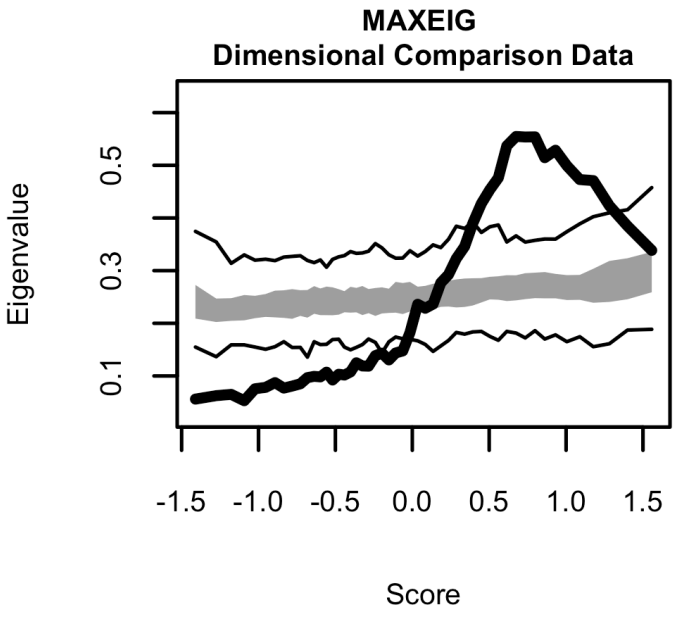
Parameter	Function and Default Value
str	The type of data to be generated. This argument has no default value; users must specify either “dim” to generate a sample of dimensional data or “cat” (or anything else) to generate a sample of categorical data.
n	Sample size. The default value is 600.
k	Number of variables. The default value is 4.
p	Taxon base rate. The default value is .5.
d	Standardized mean difference between groups. The default value is 2.
r	Correlation among variables. The default value is 0.
r.tax	Correlation among variables within the taxon. The default value is 0.
r.comp	Correlations among variables within the complement. The default value is 0.
skew	Amount of skew to be applied to variables. The default value is 0.
cuts	Number of values to use when generating ordered categorical data. The default value is 0.
seed	Random number seed; specifying the same seed enables users to generate and analyze identical data sets. The default value is 1.

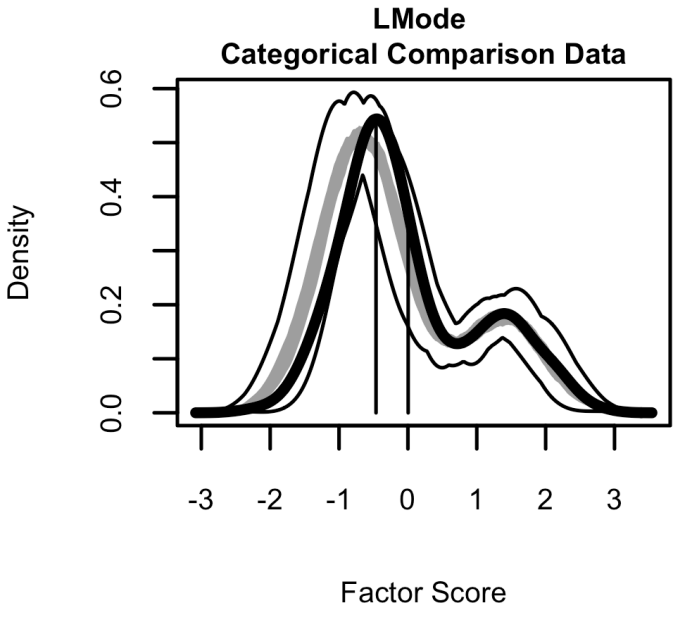
Notes. The CreateData() function allows users to create artificial datasets of known structure (categorical or dimensional), with the data parameters and default values shown here.

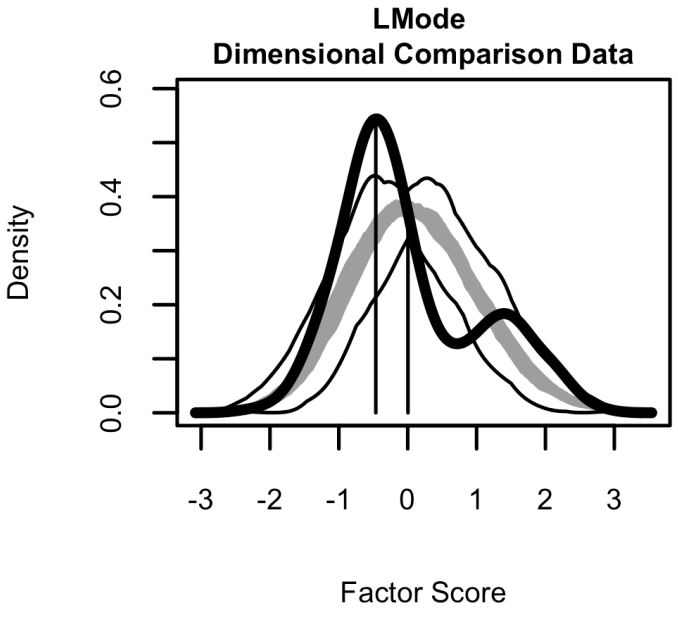

MAMBAC
Categorical Comparison Data
Mean Difference
0.4
0.8
1.2
0
10
20
30
40
50
Cut

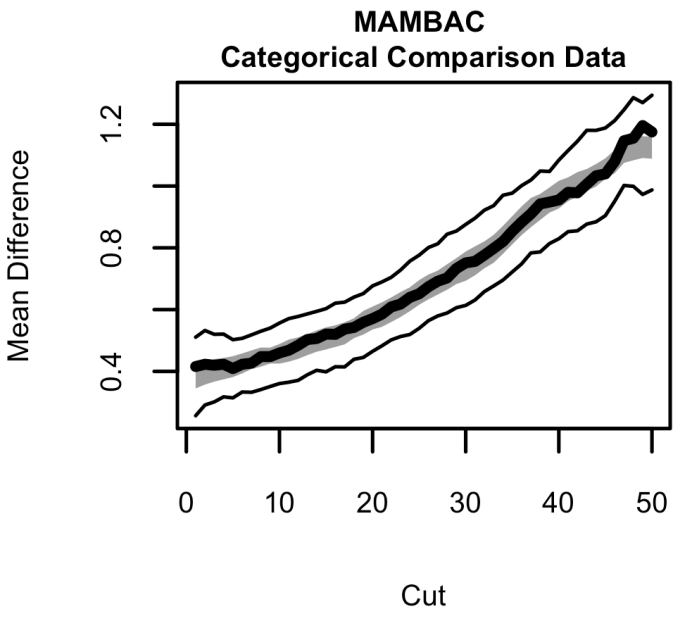

MAMBAC
Dimensional Comparison Data
Mean Difference
0.4
0.8
1.2
0
10
20
30
40
50
Cut

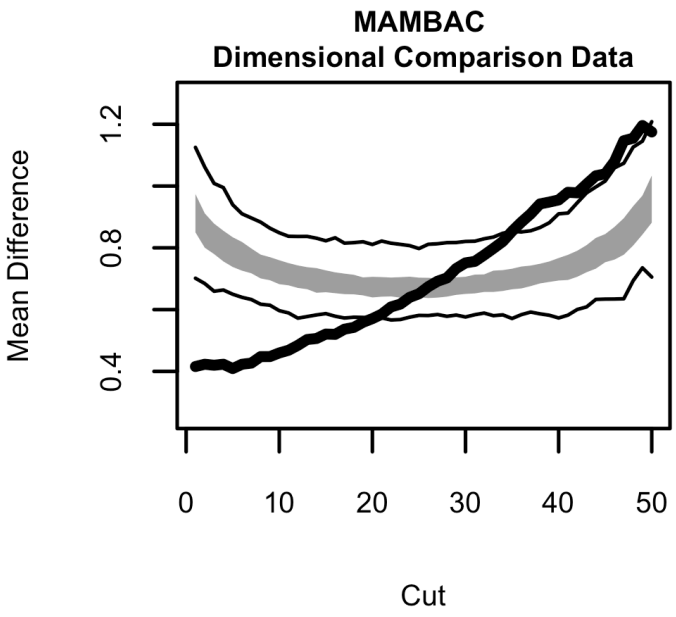

MAXEIG
Categorical Comparison Data
Eigenvalue
0.1
0.3
0.5
-1.5
-1.0
-0.5
0.0
0.5
1.0
1.5
Score

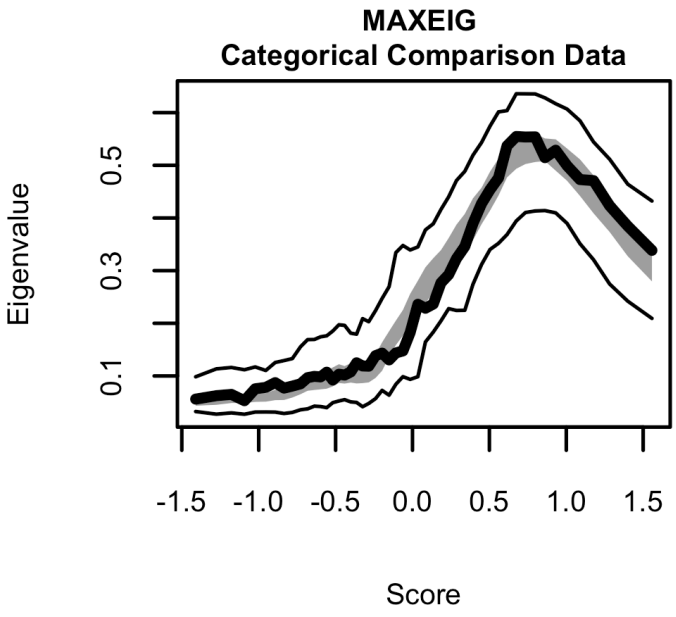

MAXEIG
Dimensional Comparison Data
Eigenvalue
0.1
0.3
0.5
-1.5
-1.0
-0.5
0.0
0.5
1.0
1.5
Score

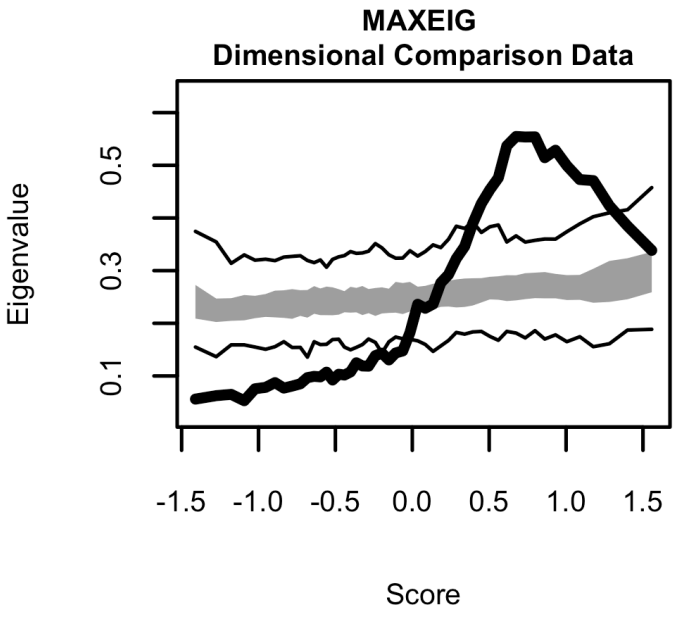

LMode
Categorical Comparison Data
Density
0.0
0.2
0.4
0.6
-3
-2
-1
0
1
2
3
Factor Score

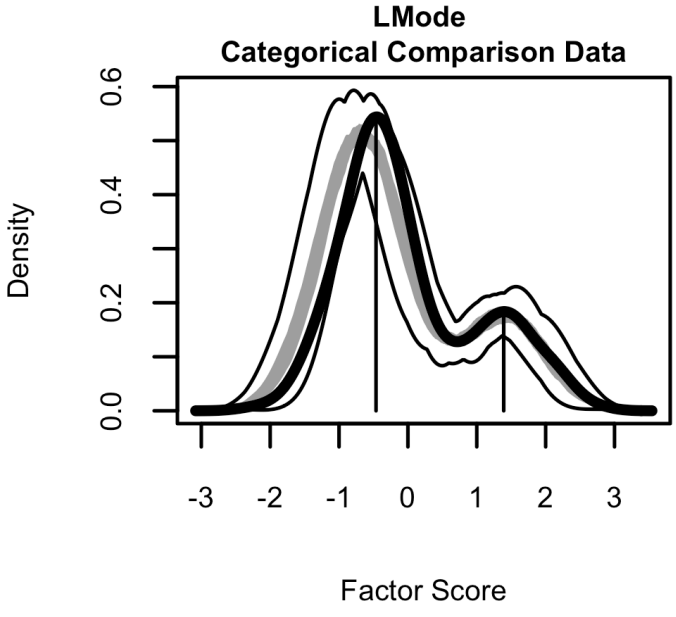

LMode
Dimensional Comparison Data
Density
0.0
0.2
0.4
0.6
-3
-2
-1
0
1
2
3
Factor Score

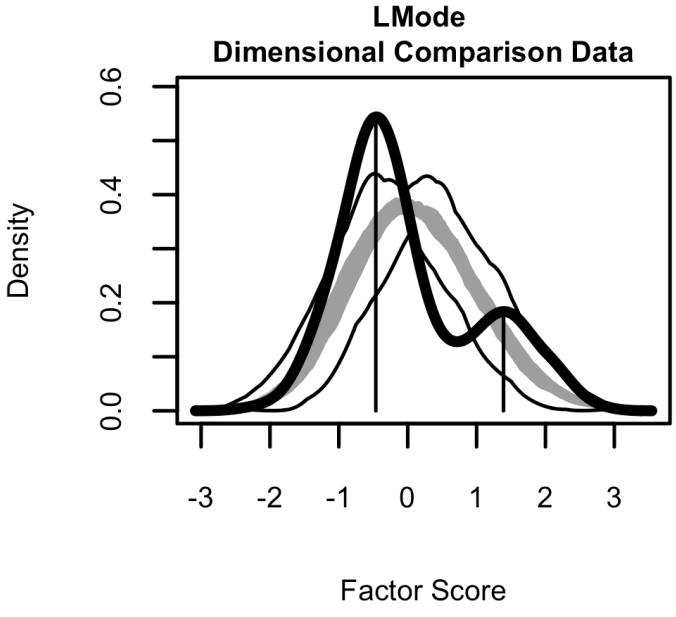

MAMBAC
Categorical Comparison Data
Mean Difference
0.4
0.8
1.2
0
10
20
30
40
50
Cut

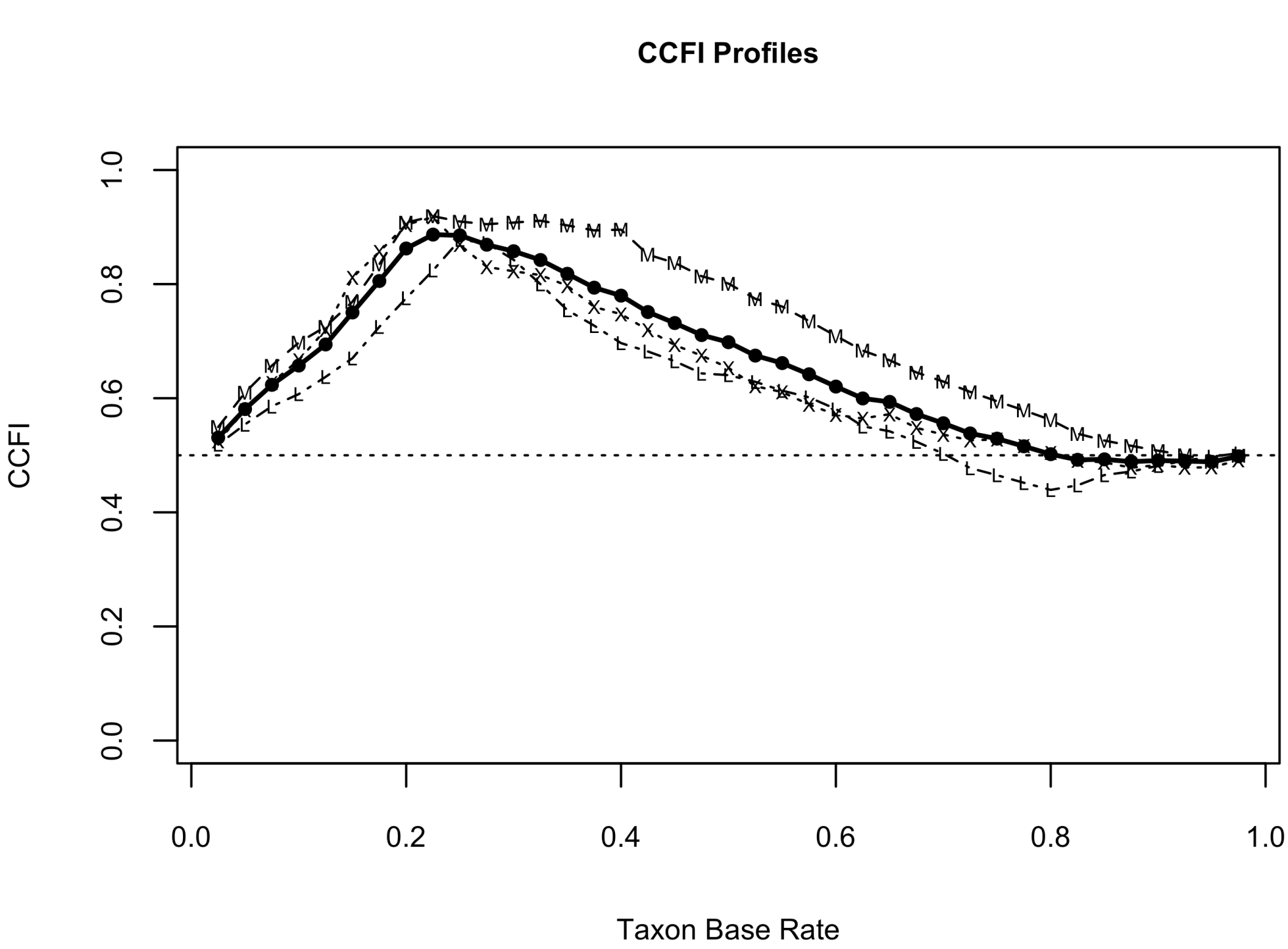

MAMBAC
Dimensional Comparison Data
Mean Difference
0.4
0.8
1.2
0
10
20
30
40
50
Cut

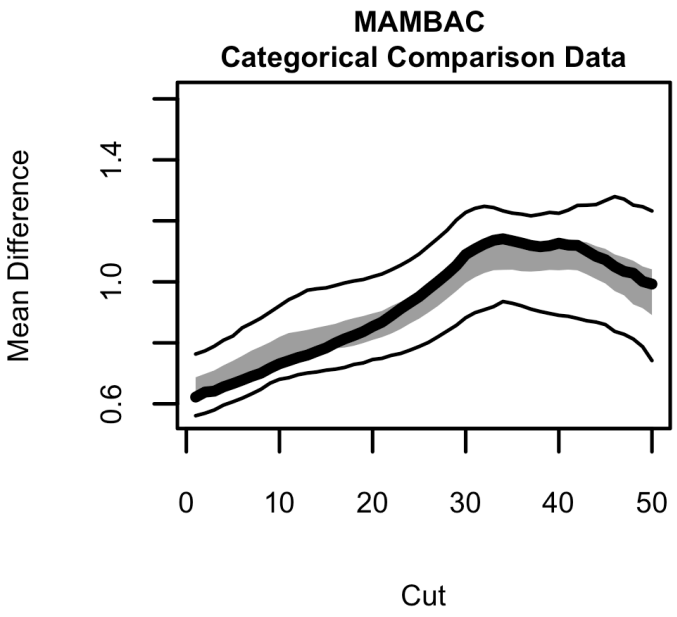

MAXEIG
Categorical Comparison Data
Eigenvalue
0.1
0.3
0.5
-1.5
-1.0
-0.5
0.0
0.5
1.0
1.5
Score

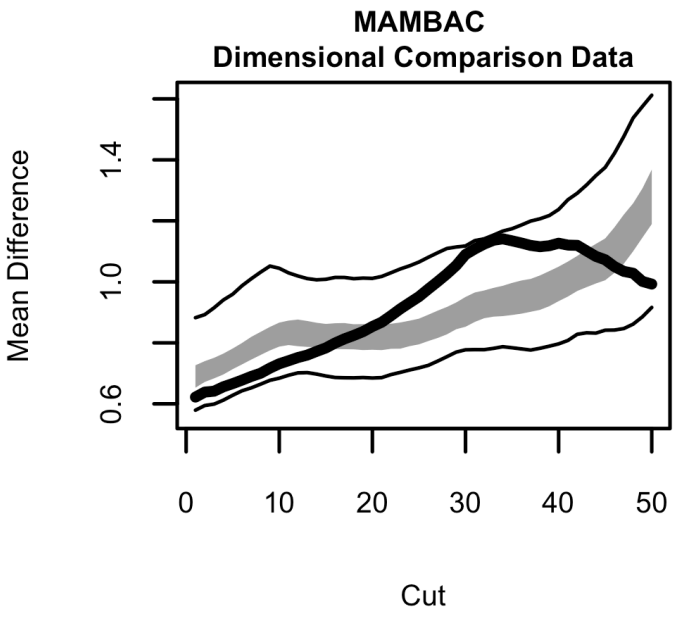

MAXEIG
Dimensional Comparison Data
Eigenvalue
0.1
0.3
0.5
-1.5
-1.0
-0.5
0.0
0.5
1.0
1.5
Score

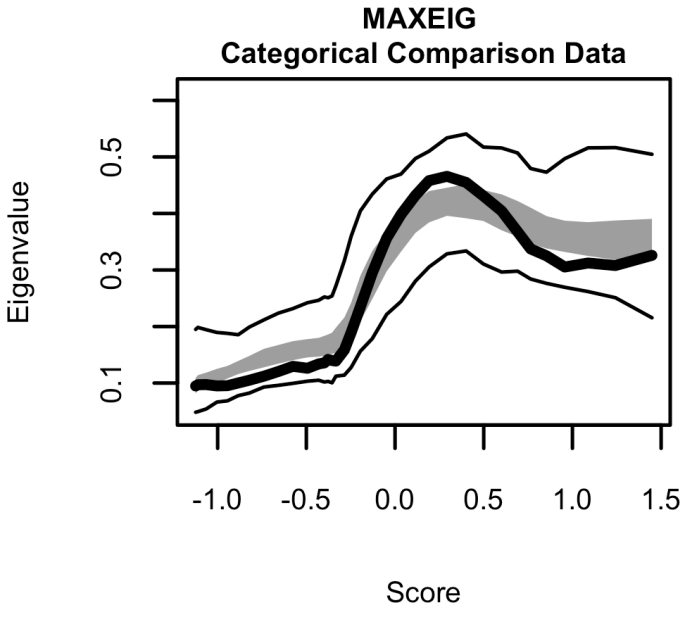

LMode
Categorical Comparison Data
Density
0.0
0.2
0.4
0.6
-3
-2
-1
0
1
2
3
Factor Score

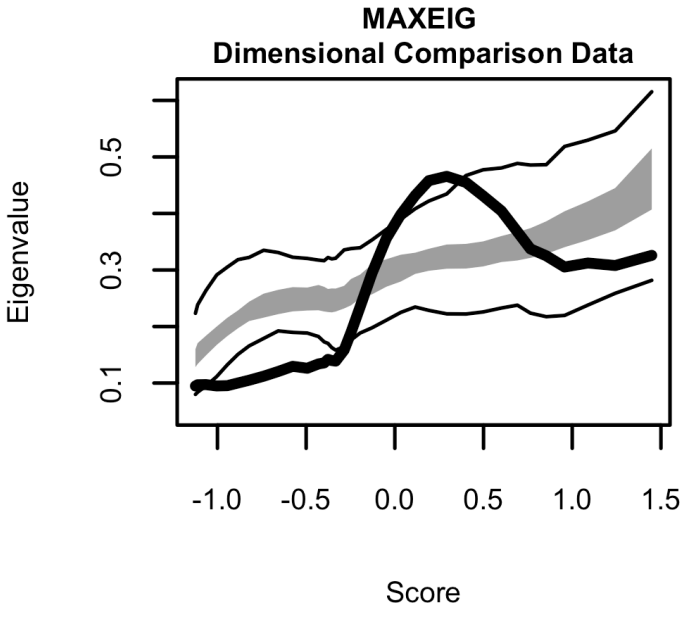

LMode
Dimensional Comparison Data
Density
0.0
0.2
0.4
0.6
-3
-2
-1
0
1
2
3
Factor Score

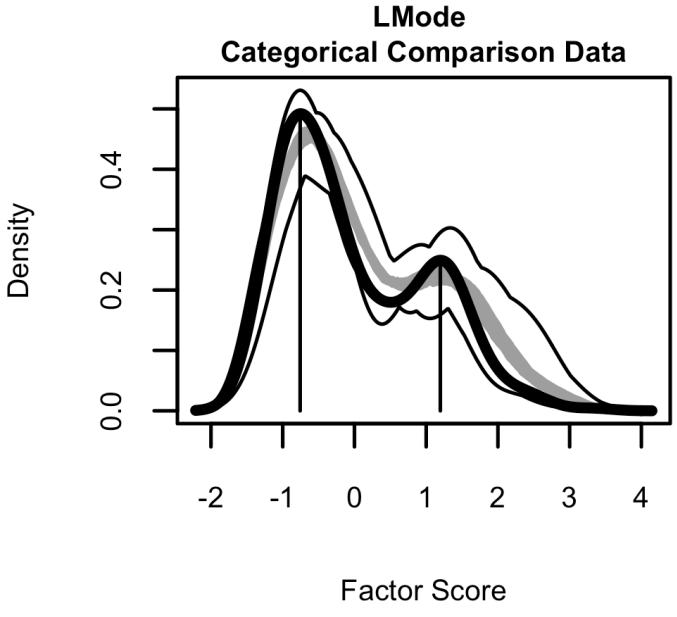

CCFI Profiles
CCFI
0.0
0.2
0.4
0.6
0.8
1.0
Taxon Base Rate

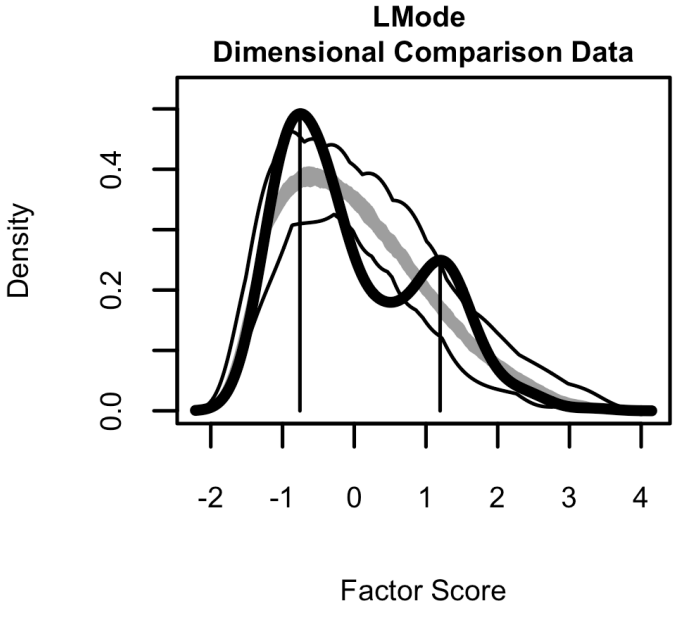

MAMBAC
Categorical Comparison Data
Mean Difference
0.6
1.0
1.4
0
10
20
30
40
50
Cut

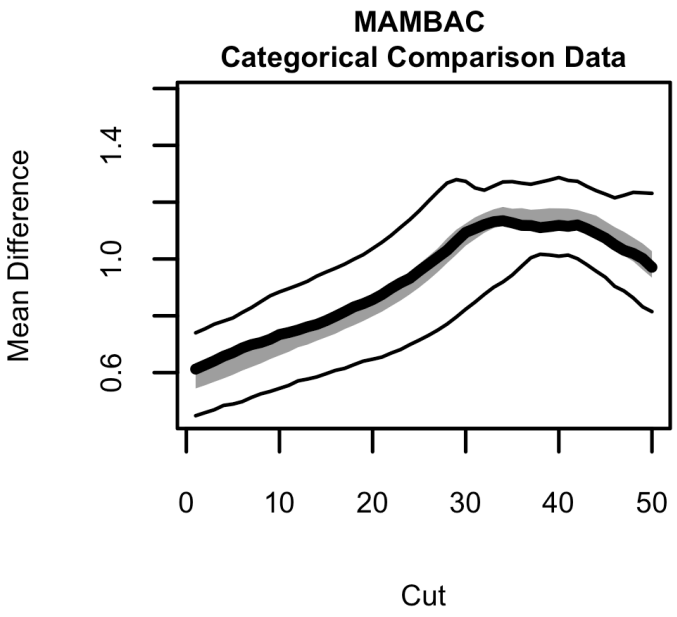

MAMBAC
Dimensional Comparison Data
Mean Difference
0.6
1.0
1.4
0
10
20
30
40
50
Cut

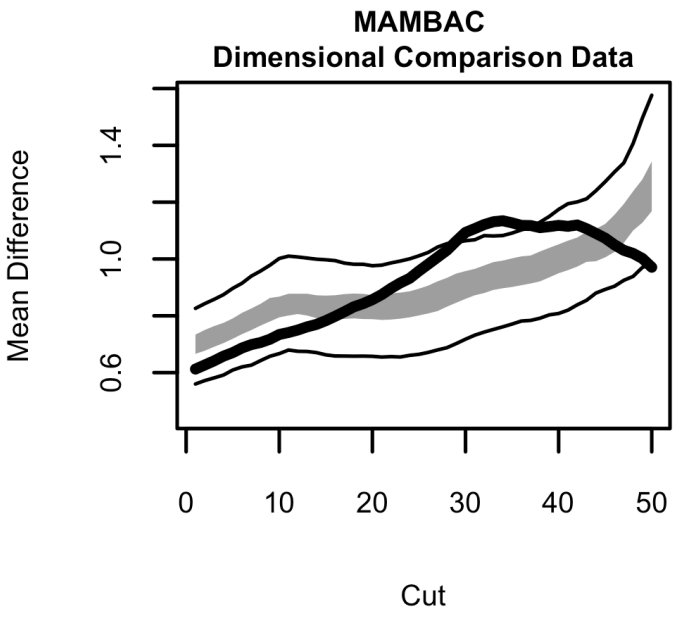

MAXEIG
Categorical Comparison Data
Eigenvalue
0.1
0.3
0.5
-1.0
-0.5
0.0
0.5
1.0
1.5
Score

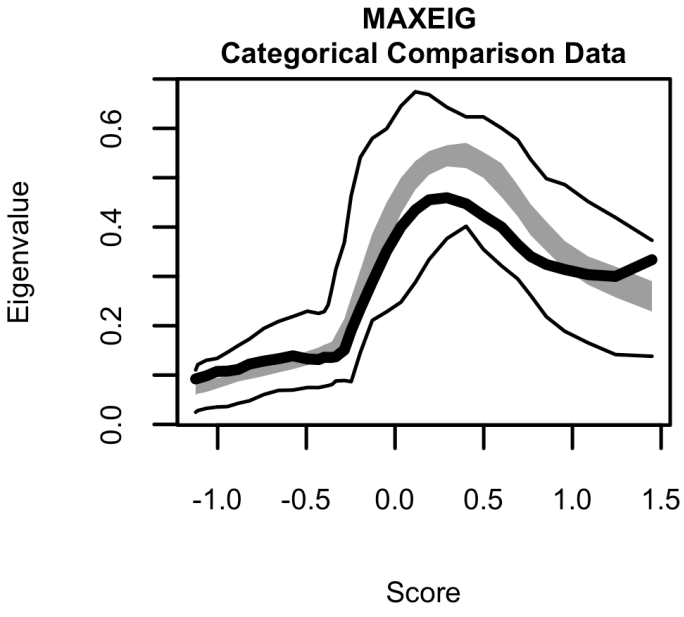

MAXEIG
Dimensional Comparison Data
Eigenvalue
0.1
0.3
0.5
-1.0
-0.5
0.0
0.5
1.0
1.5
Score

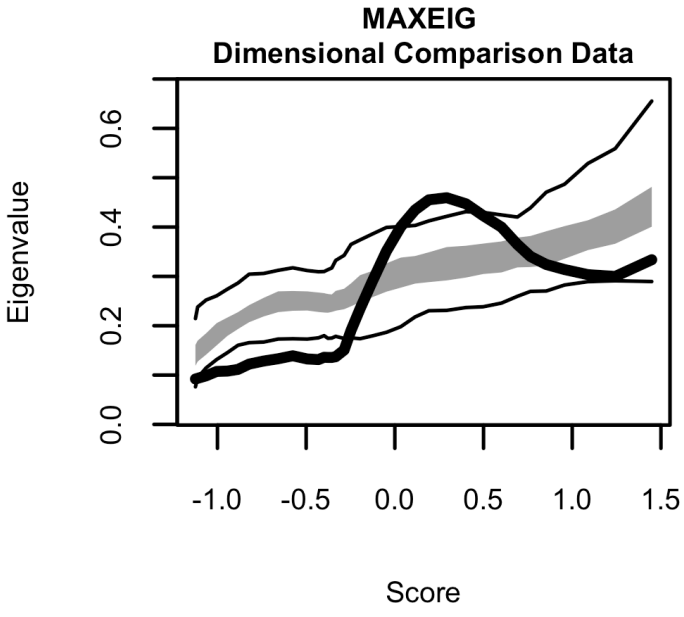

LMode
Categorical Comparison Data
Density
0.0
0.2
0.4
-2
-1
0
1
2
3
4
Factor Score

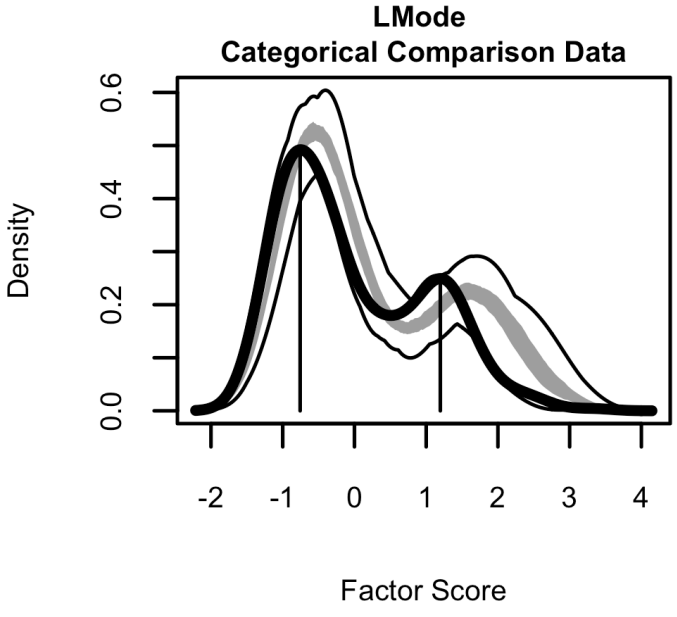

LMode
Dimensional Comparison Data
Density
0.0
0.2
0.4
-2
-1
0
1
2
3
4
Factor Score

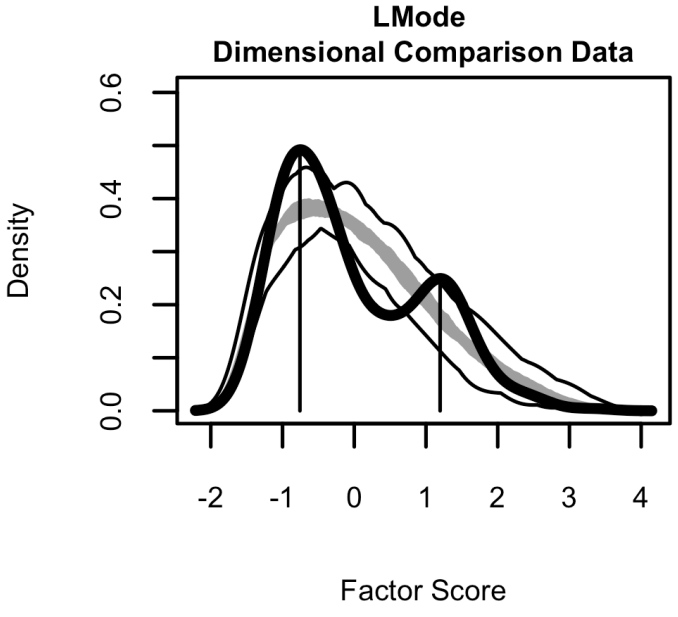

MAMBAC
Categorical Comparison Data
Mean Difference
0.6
1.0
1.4
0
10
20
30
40
50
Cut

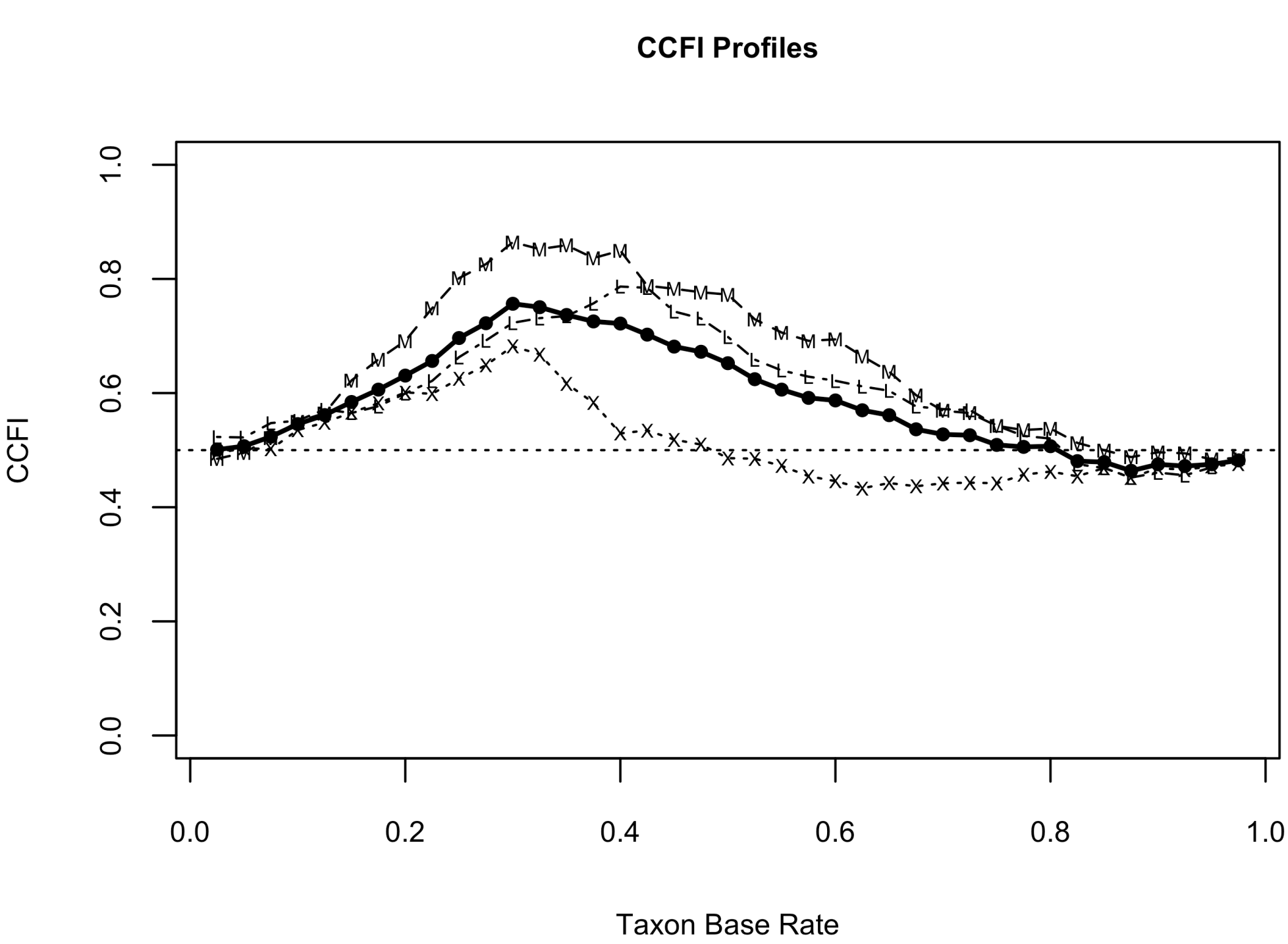

MAMBAC
Dimensional Comparison Data
Mean Difference
0.6
1.0
1.4
0
10
20
30
40
50
Cut

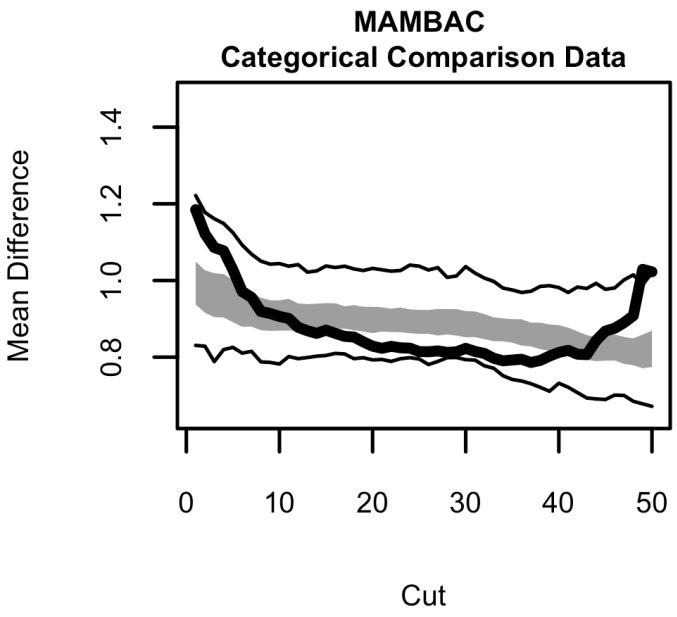

MAXEIG
Categorical Comparison Data
Eigenvalue
0.0
0.2
0.4
0.6
-1.0
-0.5
0.0
0.5
1.0
1.5
Score

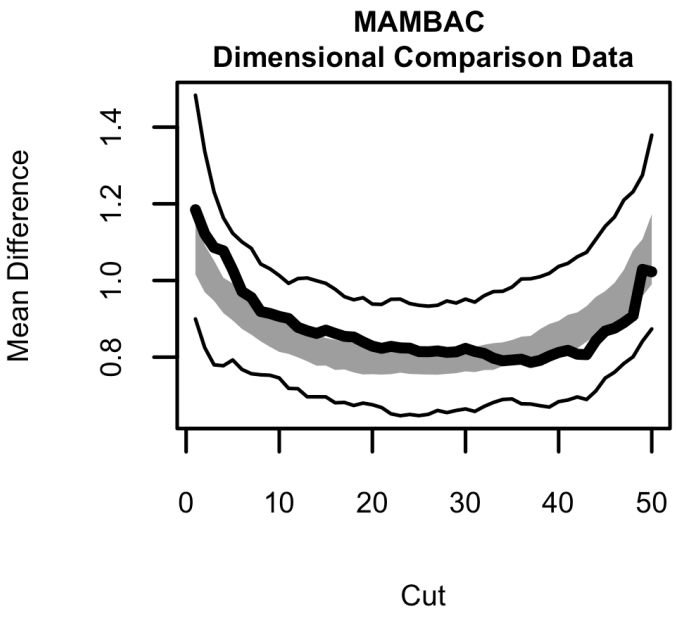

MAXEIG
Dimensional Comparison Data
Eigenvalue
0.0
0.2
0.4
0.6
-1.0
-0.5
0.0
0.5
1.0
1.5
Score

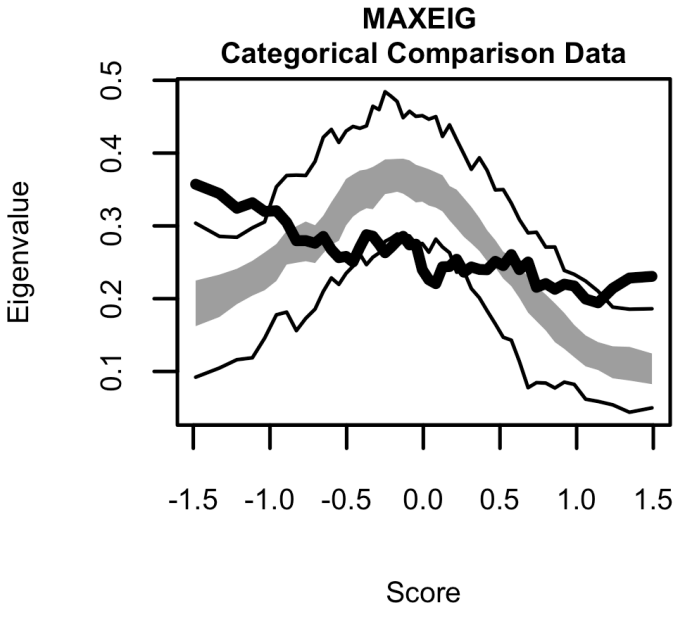

LMode
Categorical Comparison Data
Density
0.0
0.2
0.4
0.6
-2
-1
0
1
2
3
4
Factor Score

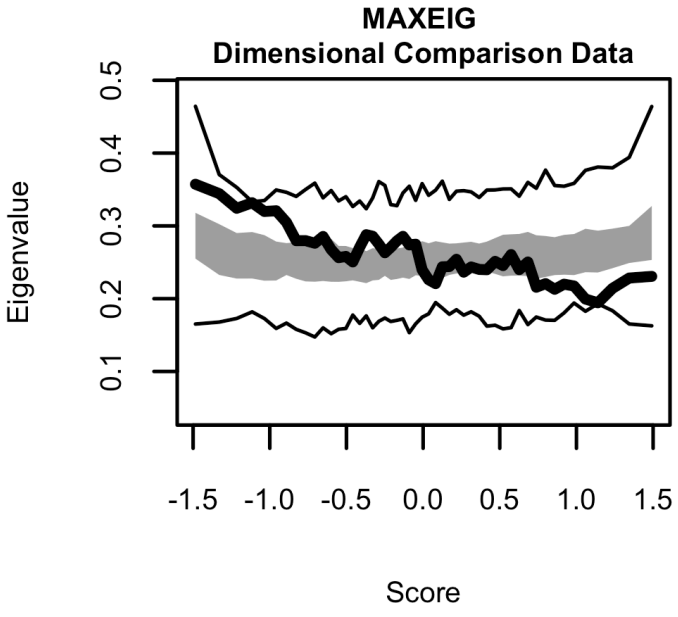

LMode
Dimensional Comparison Data
Density
0.0
0.2
0.4
0.6
-2
-1
0
1
2
3
4
Factor Score

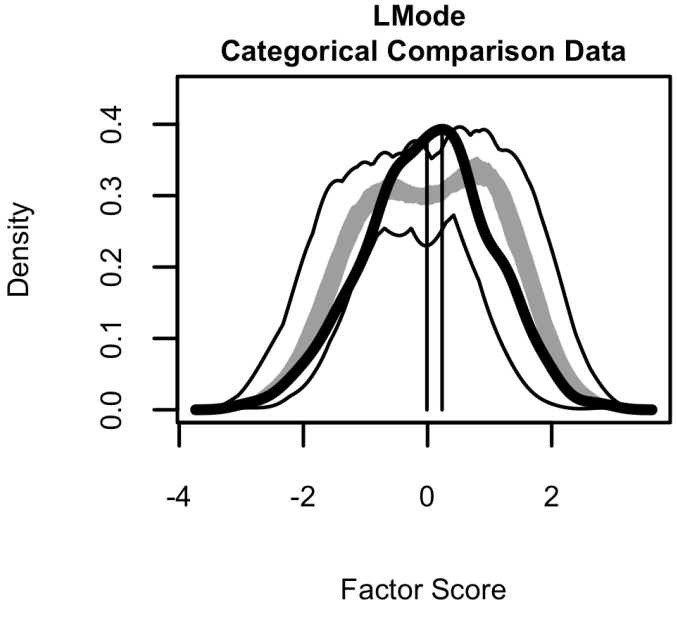

CCFI Profiles
CCFI
0.0
0.2
0.4
0.6
0.8
1.0
Taxon Base Rate

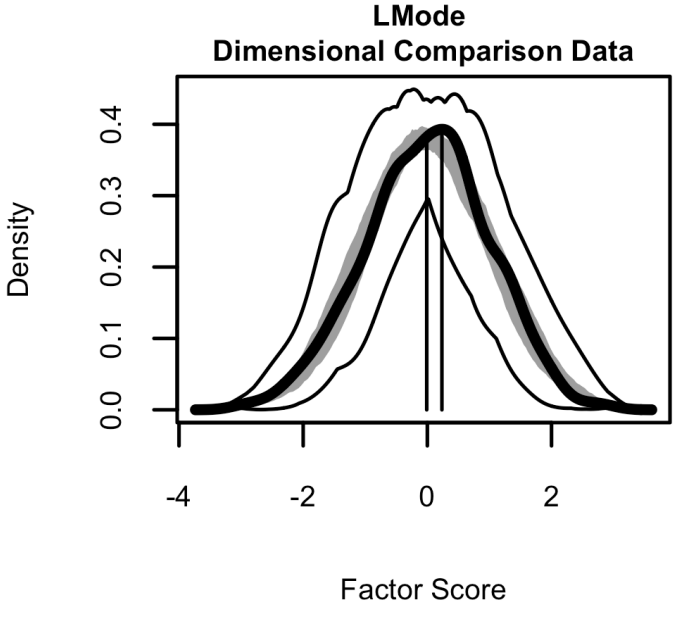

MAMBAC
Categorical Comparison Data
Mean Difference
0.8
1.0
1.2
1.4
0
10
20
30
40
50
Cut

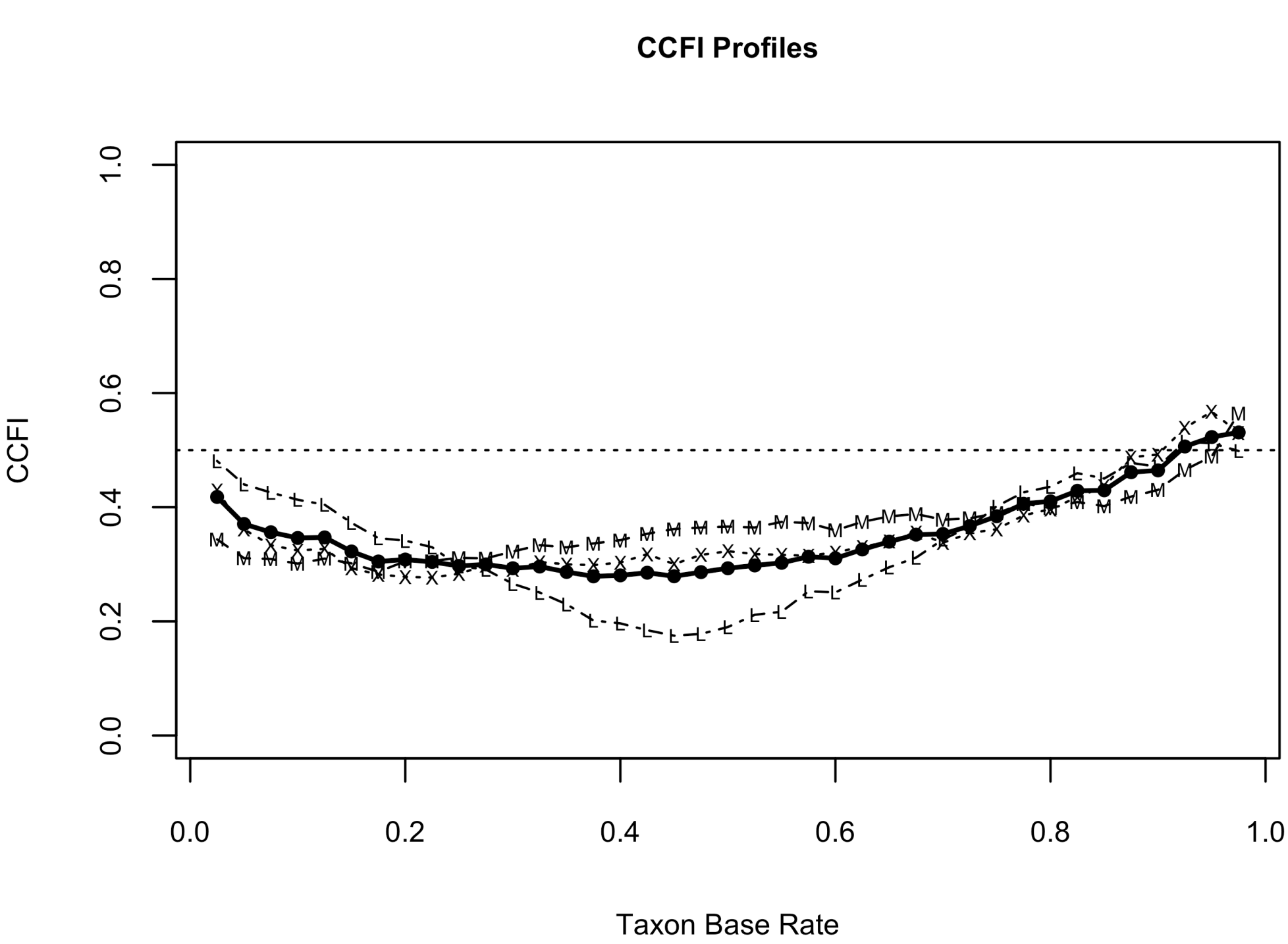

MAMBAC
Dimensional Comparison Data
Mean Difference
0.8
1.0
1.2
1.4
0
10
20
30
40
50
Cut

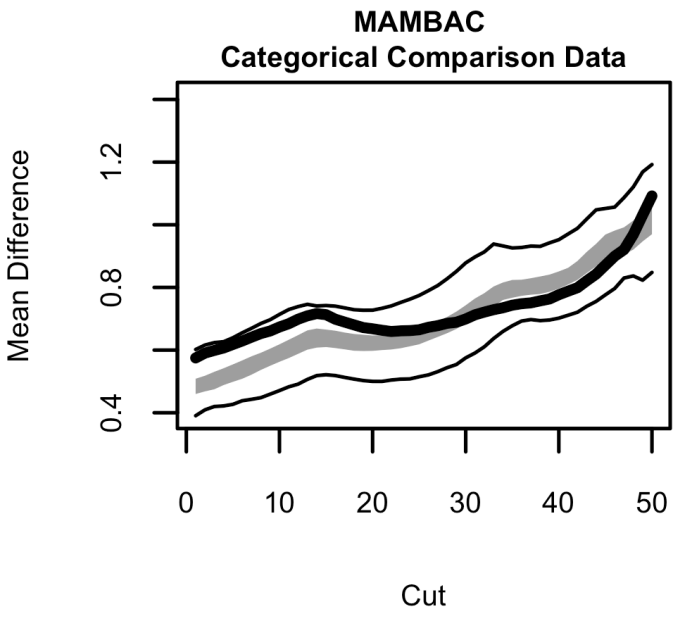

MAXEIG
Categorical Comparison Data
Eigenvalue
0.1
0.2
0.3
0.4
0.5
-1.5
-1.0
-0.5
0.0
0.5
1.0
1.5
Score

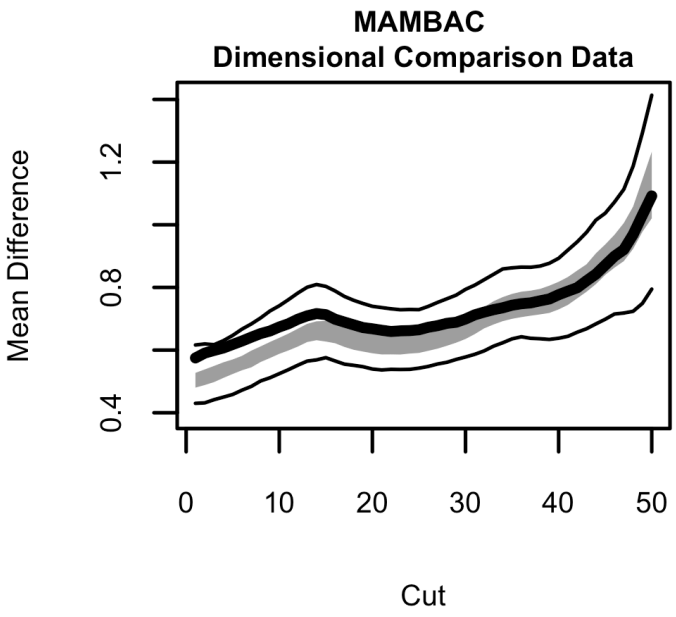

MAXEIG
Dimensional Comparison Data
Eigenvalue
0.1
0.2
0.3
0.4
0.5
-1.5
-1.0
-0.5
0.0
0.5
1.0
1.5
Score

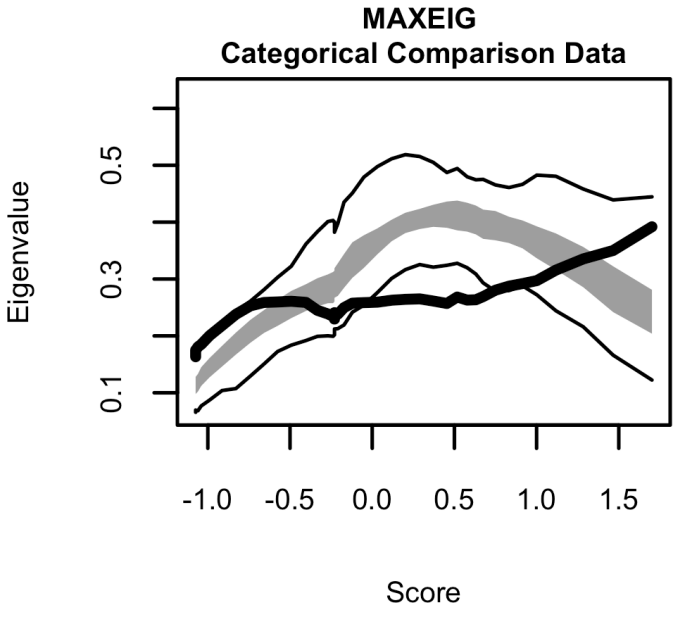

LMode
Categorical Comparison Data
Density
0.0
0.1
0.2
0.3
0.4
-4
-2
0
2
Factor Score

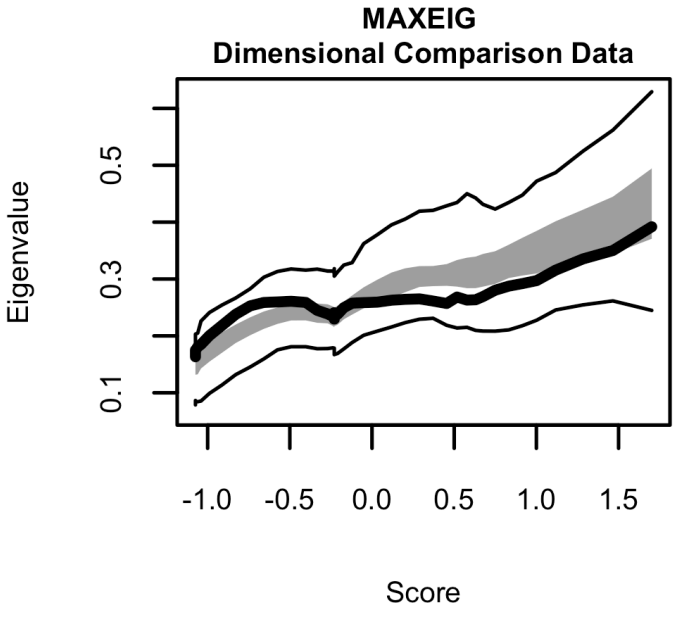

LMode
Dimensional Comparison Data
Density
0.0
0.1
0.2
0.3
0.4
-4
-2
0
2
Factor Score

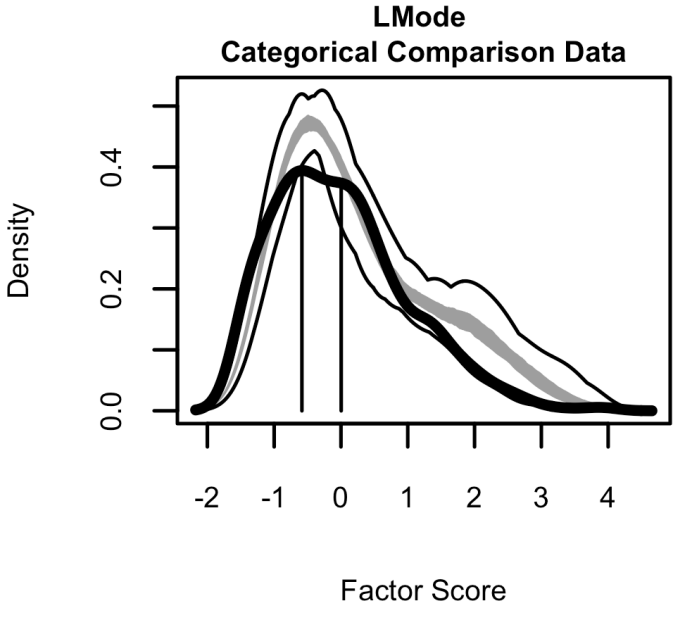

CCFI Profiles
CCFI
0.0
0.2
0.4
0.6
0.8
1.0
Taxon Base Rate

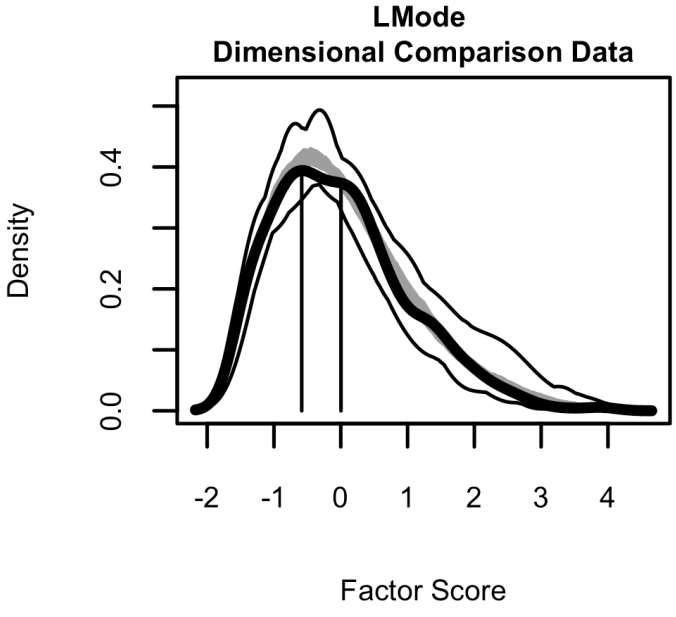

MAMBAC
Categorical Comparison Data
Mean Difference
0.4
0.8
1.2
0
10
20
30
40
50
Cut


MAMBAC
Dimensional Comparison Data
Mean Difference
0.4
0.8
1.2
0
10
20
30
40
50
Cut


MAXEIG
Categorical Comparison Data
Eigenvalue
0.1
0.3
0.5
-1.0
-0.5
0.0
0.5
1.0
1.5
Score


MAXEIG
Dimensional Comparison Data
Eigenvalue
0.1
0.3
0.5
-1.0
-0.5
0.0
0.5
1.0
1.5
Score


LMode
Categorical Comparison Data
Density
0.0
0.2
0.4
-2
-1
0
1
2
3
4
Factor Score


LMode
Dimensional Comparison Data
Density
0.0
0.2
0.4
-2
-1
0
1
2
3
4
Factor Score