Computational Statistics and Data Analysis 107 (2017) 81–91

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Deriving optimal data-analytic regimes from benchmarking studies

Lisa L. Doove [a,*], Tom F. Wilderjans [a,b], Antonio Calcagnì [a,c], Iven Van Mechelen [a]

[a] Department of Psychology, Katholieke Universiteit Leuven, Tiensestraat 102-bus 3713, Leuven, Belgium

[b] Methods and Statistics Unit, Institute of Psychology, Faculty of Social and Behavioral Sciences, Leiden University, Wassenaarseweg 52, 2333 AK Leiden, The Netherlands

[c] Department of Psychology and Cognitive Science, University of Trento, Corso Bettini, 31-38068 Rovereto, Italy

ARTICLE INFO

Article history:
Received 1 December 2015
Received in revised form 16 October 2016
Accepted 17 October 2016
Available online 27 October 2016

Keywords:
Additive clustering
Benchmarking
Classification trees
Comparison of methods
Data-analytic regimes
Supervised learning

ABSTRACT

In benchmarking studies with simulated data sets in which two or more statistical methods are compared, over and above the search of a universally winning method, one may investigate how the winning method may vary over patterns of characteristics of the data or the data-generating mechanism. Interestingly, this problem bears strong formal similarities to the problem of looking for optimal treatment regimes in biostatistics when two or more treatment alternatives are available for the same medical problem or disease. It is outlined how optimal data-analytic regimes, that is to say, rules for optimally calling in statistical methods, can be derived from benchmarking studies with simulated data by means of supervised classification methods (e.g., classification trees). The approach is illustrated by means of analyses of data from a benchmarking study to compare two different algorithms for the estimation of a two-mode additive clustering model.

© 2016 Elsevier B.V. All rights reserved.

1. Introduction

For any statistical or data-analytic process, many choice options are available on the level of the method of preprocessing of the input, of the actual method of data analysis, and of the method of post-processing the output. An obvious question in such cases reads which choice alternatives are optimal in some sense. This comes down to a question of comparative evaluation, the goal of which being to identify the optimal method. This question is crucial for practitioners who may wish to make good choices in statistical practice. A typical setting to study such a comparative evaluation is that of benchmark studies, in which simulated or empirical data are used to evaluate several methods under study. A possible result of such studies is the identification of a universal winning method (i.e., a gold standard).

However, often the winning method varies across types of data sets, with the data sets in question being typified in terms of structural characteristics (e.g., data size) or of characteristics of the underlying data-generating mechanism (e.g., noise structure). As an example, Schepers et al. (2006) performed a benchmarking study with simulated data sets to investigate the sensitivity of different partitioning algorithms to local optima. In this study, at first sight, there appeared to be a universal winning method with regard to the criterion "proportion of multistart runs that return proxy of global optimum". Yet,

* Corresponding author.
E-mail addresses: lisa.doove@kuleuven.be (L.L. Doove), t.f.wilderjans@fsw.leidenuniv.nl (T.F. Wilderjans), antonio.calcagni@unitn.it (A. Calcagnì), iven.vanmechelen@kuleuven.be (I. Van Mechelen).

http://dx.doi.org/10.1016/j.csda.2016.10.016
0167-9473/© 2016 Elsevier B.V. All rights reserved.

subsequently the performance order of the algorithms appeared to change in case of a different data-generating mechanism for the residuals (viz., one that implied between-residual dependencies). When the winning method varies across types of data sets, the question "Which method is best?" becomes meaningless and should be replaced by "Which method is best for which type of data set?" Moreover, the answer to the latter question may also depend on the evaluation criterion that one is interested in. We will address this issue by looking for rules that indicate which method one should optimally use for a data set at hand, given some evaluation criterion, and given characteristics of the data set under study and of its underlying data-generating mechanism. For reasons that will become clear below, we will further call such rules "optimal data-analytic regimes".

The identification of optimal data-analytic regimes is quite a challenge as it goes with two major problems. The first one pertains to the fact that data sets under study may differ from one another on a huge number of characteristics, with the optimal data-analytic method possibly depending on complex patterns of them. Moreover, often precise hypotheses on the specific relation between patterns of characteristics and method optimality are lacking; this implies that the types of data sets across which the winning method varies are to be induced during the actual analysis for identifying optimal data-analytic regimes. The second problem pertains to the risk of sizeable inferential errors that the identification of optimal data-analytic regimes may be prone to. On the one hand, these include the possibility that types of data sets across which the winning method varies are not detected. On the other hand, one should also beware of erroneously ending up with a nontrivial data-analytic regime while in fact there is a universal winning method.

As a possible solution for these problems, in this paper, we propose a novel methodology to identify optimal data-analytic regimes. This novel methodology is based in part on principles borrowed from the domain of optimal treatment regime estimation in biostatistics, that is, the identification of rules that indicate how the optimal treatment alternative may depend on pretreatment or baseline characteristics of the target patients under study (see, e.g., Zhang et al., 2012). We primarily focus on optimal data-analytic regimes derived from benchmarking studies with simulated data in which two methods are compared. Extensions to a comparison of more than two methods and to studies with empirical data, which are not fully straightforward, will be addressed in the discussion.

The remainder of this paper has been structured as follows. In Section 2, we define a conceptual framework in which we formalize the problem at hand, and we introduce the novel methodology. We illustrate with data from a benchmarking study to compare two algorithms for estimating a two-mode additive biclustering model in Section 3. Discussion points and concluding remarks are given in Section 4.

2. Framework and methodology

2.1. Benchmarking study with simulated data

In this paper we focus on benchmarking studies with data that have been simulated according to a full factorial design. The factors $\mathbf{X}$ of this design pertain to characteristics of the data and the data-generating mechanism (e.g., data size, nature and complexity of data-generating mechanism, noise level, and noise structure). For each combination of levels of the factors under study, a number r of simulated data sets is generated. Subsequently, each simulated data set is analyzed with the two methods that are to be compared. These may be methods of pre-processing of input, actual methods of data analysis, or methods of post-processing of output (including model selection heuristics). Subsequently, some outcome of interest, Y, is evaluated. Variable Y may pertain to optimization performance, technical quality (e.g., computational cost, stability, replicability), recovery performance (i.e., recovery of aspects of the underlying true model, such as the true level of complexity of this model or true parameter values), or performance with regard to possible other goals of the data analysis (e.g., predictive quality, relation with external criterion variables). All this implies that the data for the optimal data-analytic regime estimation are the values of an outcome variable Y for data from a mixed factorial design with as between-factors characteristics of the data and the data-generating mechanism ($\mathbf{X}$), with a binary within-factor A pertaining to method (taking values 0 and 1), and with r replications per cell of the design.

2.2. Data-analytic regimes

Given data from a benchmarking study as outlined above, a data-analytic regime is a function g that maps the value set of characteristics of the data and the data-generating mechanism $\mathbf{X}$ to the values of A. This function formalizes the rule that a data set with pattern of data characteristics $\mathbf{X}$ is to be analyzed with Method 0 if $g(\mathbf{X}) = 0$ and with Method 1 if $g(\mathbf{X}) = 1$.

Let Y^a denote the outcome of a data set subjected to Method a. Then, the outcome that would result from assigning a data-analytic method to a randomly chosen data set according to g is given by

$$Y^{g(\mathbf{X})} = Y^1 g(\mathbf{X}) + Y^0[1 - g(\mathbf{X})].$$

Assuming, without loss of generality, that larger values of Y are more desirable, the optimal data-analytic regime then is the one leading to the largest expected outcome $E(Y^{g(\mathbf{X})})$ when applied to the population of data sets under study. In what follows, we will focus on a pre-specified class of data-analytic regimes denoted by $\mathcal{G}$. Given $\mathcal{G}$, we may define the optimal data-analytic regime, $g^{\text{opt}}(\mathbf{X})$, as the one leading to the largest value of $E(Y^{g(\mathbf{X})})$ among all $g \in \mathcal{G}$, that is to say,

$$g^{\text{opt}}(\mathbf{X}) = \arg\max_{g \in \mathcal{G}} E(Y^{g(\mathbf{X})}).$$

Table 1
Four types of characteristics of the data and the data-generating mechanism in benchmarking studies with simulated data.

Manipulated in simulation study	Known in data-analytic practice	
	No	Yes
Yes	(a)	(b)
No	(c)	(d)

In estimating the optimal data-analytic regime given data from a benchmarking study with simulated data, different types of characteristics of the data and the data-generating mechanism can be considered for $\mathbf{X}$. The most obvious type consists of characteristics that are manipulated in the benchmarking study. At this point, however, two complications may arise: First, besides the manipulated characteristics, other sources of variation may be relevant for the optimal data-analytic regime. As an example, the skewness of some variables in the simulated data might not have been directly manipulated by the investigator, whereas it may differ across simulated data sets and influence the performance of data-analytic methods. Otherwise, unmanipulated characteristics can either fluctuate at random across the simulated data sets, or they can be manipulated indirectly through the characteristics of the data and data-generating mechanism that have been varied experimentally by the researcher. Second, characteristics that are known in a Monte Carlo simulation setting may be unknown in data-analytic practice. As an example, characteristics such as noise level and noise structure are often experimentally varied in simulation studies, yet, they are typically unknown for empirical data sets outside a simulation context. If estimated optimal data-analytic regimes would rely on such characteristics of the data and the data-generating mechanism that are unknown in data-analytic practice, the regimes in question could not be readily used by practitioners to determine the optimal method for a data set at hand.

Summing up, by combining the two facets outlined above, four types of characteristics of the data and the data-generating mechanism may be distinguished (see Table 1). In the estimation of optimal data-analytic regimes, there is a clear preference for regimes that rely on characteristics that are known to the practitioner outside a Monte Carlo context (cells (b) and (d) of Table 1). Besides, from the viewpoint of inferential certainty, there may be some preference for characteristics that have been manipulated (cells (a) and (b)) over characteristics that have not been manipulated (cells (c) and (d)).

2.3. Estimation of data-analytic regimes

To obtain an estimator $\hat{g}^{\text{opt}}$ for g^{opt}, we will rely in part on earlier work on the identification of optimal treatment regimes in the field of biostatistics (Zhang et al., 2012). We will first transform the problem of estimating optimal data-analytic regimes into a weighted classification problem, which will allow us to use existing classification techniques to estimate optimal data-analytic regimes. Subsequently, we will propose a stepwise procedure to estimate optimal data-analytic regimes that are useful in data-analytic practice (i.e., that involve characteristics in cells (b) and (d) of Table 1).

To transform the problem of estimating optimal data-analytic regimes into a weighted classification problem, we observe that

$$\begin{aligned} E(Y^{g(\mathbf{X})}) &= E\{Y^1 g(\mathbf{X}) + Y^0[1 - g(\mathbf{X})]\} \\ &= E\{g(\mathbf{X})[Y^1 - Y^0] + Y^0\} \\ &= E[g(\mathbf{X})C(\mathbf{X}) + Y^0], \end{aligned} \tag{1}$$

where $C(\mathbf{X}) = Y^1 - Y^0$ is a so-called contrast function that denotes for each data set the difference in measured outcome between Method 1 and Method 0. It follows that $g^{\text{opt}}(\mathbf{X}) = \arg\max_{g\in\mathcal{G}} E(Y^{g(\mathbf{X})}) = \arg\max_{g\in\mathcal{G}} E[g(\mathbf{X})C(\mathbf{X})]$. Next, because $C(\mathbf{X}) = I[C(\mathbf{X}) > 0]|C(\mathbf{X})| - I[C(\mathbf{X}) \leq 0]|C(\mathbf{X})|$, $g(\mathbf{X})C(\mathbf{X})$ can be rewritten as

$$\begin{aligned} g(\mathbf{X})C(\mathbf{X}) &= g(\mathbf{X})I[C(\mathbf{X}) > 0]|C(\mathbf{X})| - g(\mathbf{X})I[C(\mathbf{X}) \leq 0]|C(\mathbf{X})| \\ &= I[C(\mathbf{X}) > 0]|C(\mathbf{X})| - |C(\mathbf{X})|\{[1 - g(\mathbf{X})]I[C(\mathbf{X}) > 0] + g(\mathbf{X})I[C(\mathbf{X}) \leq 0]\}. \end{aligned}$$

As $g(\mathbf{X})$ takes values $\{0, 1\}$,

$$[1 - g(\mathbf{X})]I[C(\mathbf{X}) > 0] + g(\mathbf{X})I[C(\mathbf{X}) \leq 0] = \{I[C(\mathbf{X}) > 0] - g(\mathbf{X})\}^2.$$

Combining these results, $g(\mathbf{X})C(\mathbf{X})$ can be rewritten as

$$g(\mathbf{X})C(\mathbf{X}) = I[C(\mathbf{X}) > 0]|C(\mathbf{X})| - |C(\mathbf{X})|\{I[C(\mathbf{X}) > 0] - g(\mathbf{X})\}^2.$$

Therefore, the optimal data-analytic regime can be estimated as

$$\begin{aligned} g^{\text{opt}}(\mathbf{X}) &= \arg\max_{g\in\mathcal{G}} E[g(\mathbf{X})C(\mathbf{X})] \\ &= \arg\max_{g\in\mathcal{G}} \left[E\{I\,[C(\mathbf{X}) > 0]\,|C(\mathbf{X})|\} - E\left(|C(\mathbf{X})|\{I[C(\mathbf{X}) > 0] - g(\mathbf{X})\}^2\right)\right] \\ &= \arg\min_{g\in\mathcal{G}} E\left(|C(\mathbf{X})|\{I\,[C(\mathbf{X}) > 0] - g(\mathbf{X})\}^2\right). \end{aligned} \tag{2}$$

At this point, $Z = I[C(\mathbf{X}) > 0]$ can be considered a known classification variable that defines two classes of data sets: (a) the class $Z = 1$ which comprises those data sets for which the outcome is higher under Method 1 ($Y^1 > Y^0$), and that should therefore be preferably analyzed with this method, and (b) the class $Z = 0$ that comprises those data sets with $Y^0 > Y^1$, and that should therefore be preferably analyzed with Method 0. Each data set is also given a weight $W = |C(\mathbf{X})|$, which represents the loss that would be incurred if the data set were misclassified.

The above implies that the information contained in $C(\mathbf{X})$ is separated into two parts: the class label Z, containing the information about the sign of $C(\mathbf{X})$; and the weight W, containing the information about the magnitude of $C(\mathbf{X})$. The estimation of the optimal data-analytic regime g^{opt} via (2) then comes down to

$$\hat{g}^{\text{opt}}(\mathbf{X}) = \arg\min_{g \in \mathcal{G}} \sum_{i=1}^{N} W_i [Z_i - g(\mathbf{X}_i)]^2,$$

which is a weighted classification problem. This problem can then be solved by means of a broad range of existing supervised classification methods (e.g., kernel methods, support vector machines, neural networks, and classification trees).

Importantly, in a benchmarking setting, a single data set can always be subjected to the different method alternatives (which implies that method can act as a within-subject or repeated measures variable). As a result, the contrast function $C(\mathbf{X})$ (and hence the class label Z and weight W) for each data set is fully known. This is different from the setting of optimal treatment regime estimation, where the application of different treatment alternatives typically is a between-subject variable, which implies that the contrast values $C(\mathbf{X})$ are inherently unknown and are therefore to be estimated from the data (Zhang et al., 2012).

2.4. Procedure to construct workable data-analytic regimes

Taking into account the four types of characteristics of the data and the data-generating mechanism as summarized in Table 1, in order to arrive at an estimated optimal data-analytic regime that is workable in data-analytic practice, we propose the following stepwise procedure:

1. **Estimate an optimal data-analytic regime using all manipulated characteristics**
   The first step is to estimate an optimal data-analytic regime $\hat{g}^{\text{opt}}(\mathbf{X})$ using all manipulated characteristics of the data and the data-generating mechanism (cells (a) and (b) of Table 1). Optionally, one may include unmanipulated characteristics, either for exploratory purposes, or because one theoretically expects particularly such characteristics to influence the performance of the data-analytic alternatives.
2. **Estimate proxies for the relevant characteristics that are unknown in data-analytic practice, and estimate a second optimal data-analytic regime using only known characteristics**
   As mentioned before, the estimated optimal data-analytic regime resulting from Step 1 may possibly involve characteristics of the data and the data-generating mechanism that are unknown in data-analytic practice (cell (a)). As a result, the resulting data-analytic regime may not be workable for practitioners to determine the optimal method for data sets in empirical practice. The second step is to estimate so-called proxies for the characteristics that appeared to be relevant in the estimated optimal data-analytic regime resulting from Step 1, but that are unknown in data-analytic practice (cell (a) of Table 1). These proxies are known (and typically unmanipulated) characteristics of the data (cell (d)) that can be considered reasonable estimates of the unknown characteristics on the basis of the previously derived optimal data-analytic regimes. As an example, the noise level in the data could be estimated on the basis of replicates (if present in the data) or on the basis of some model estimation procedure. After replacing the relevant characteristics in cell (a) by known reasonable estimates in cell (d), one should then re-estimate the optimal data-analytic regime while using only characteristics that are known in data-analytic practice (cells (b) and (d)). Optionally, one could force the classification algorithm to use the proxies in the estimation process. In case the re-estimated data-analytic regime would not be similar to the original data-analytic regime derived in Step 1, one may wish to look for other proxies of the relevant characteristics in cell (a).
3. **Testing the resulting regimes (optional)**
   If the estimation results in a nontrivial data-analytic regime, this suggests the presence of a qualitative or disordinal data set type by method interaction, meaning that for some types of data sets Method 1 outperforms Method 0, whereas for other types of data sets the reverse holds true. Rather than considering at this point tests for specific data types (which are prone to multiple testing problems — see, e.g., Lagakos, 2006), one may wish to check this by a global test of the null hypothesis of absence of a qualitative or disordinal interaction. For this purpose, one may make use of suitable test statistics such as the one proposed by Gail and Simon (1985). Two issues should be taken into account at this point: Firstly, the test in question would be a pseudo test as the set of types of data sets across which the optimal method might differ has been induced by estimating a data-analytic regime from the same data as the ones used for the hypothesis testing (which inflates the probability of false positives). Secondly, tests of the hypothesis that an interaction is disordinal generally lack statistical power (Brookes et al., 2001; Rothwell, 2005; Shaffer, 1991). In the context of benchmarking studies with simulated data sets, however, the latter should not be expected to be a problem as simulation studies typically allow for large sample sizes.

4. **Cross-validate the results in replication studies (optional)**
   The proposed optimal regime estimation procedure can be considered an exploratory post-hoc method. Hence, it should be looked at as a hypothesis-generating device, the results of which are to be cross-validated in replication studies. In such replication studies, variables that were not experimentally manipulated in the original benchmarking study (cells (c) and especially (d) of Table 1), and that showed up in the basis of the regimes resulting from Step 1, should be experimentally manipulated.

As a final note, it is important to emphasize that optimal data-analytic regimes should preferably go with sound theoretical justifications. For this purpose, it is also important to include in the design of simulation research for benchmarking characteristics that can theoretically be expected to influence the performance of the data-analytic methods under study.

3. Illustrative application

3.1. Data

We applied the newly proposed method for optimal data-analytic regime estimation to data from a benchmarking study by Wilderjans et al. (2013b) on additive biclustering. The authors considered (real-valued) object by variable data sets, for which they wanted to estimate an additive biclustering model (Baier et al., 1997; DeSarbo, 1982). This model implies: (a) (possibly overlapping) biclusters (i.e., Cartesian products of object and variable clusters), (b) a (real-valued) weight for each bicluster, and (c) that the reconstructed value for object i on variable j equals the sum of the weights of all biclusters to which (i, j) belongs. In the associated data analysis, the additive biclustering model is fitted to a given data set by means of minimizing a least squares loss function.

More formally speaking, additive biclustering with K object clusters and L variable clusters implies that a real-valued $I \times J$ object by variable data matrix $\mathbf{X}$ is approximated by an $I \times J$ real-valued model matrix $\mathbf{M}$ that can be decomposed into a binary (0/1) object cluster membership matrix $\mathbf{A}$ and a binary variable cluster membership matrix $\mathbf{B}$ of sizes $I \times K$ and $J \times L$, respectively, and a real-valued $K \times L$ weight matrix $\mathbf{V}$:

$$\mathbf{M} = \mathbf{AVB}'.$$

Given an $I \times J$ data matrix $\mathbf{X}$, and numbers of underlying object and variable clusters K and L, the aim of an additive biclustering analysis is to estimate binary matrices $\mathbf{A}$ and $\mathbf{B}$ and a real-valued matrix $\mathbf{V}$ such that the least squares loss function

$$F(\mathbf{A}, \mathbf{B}, \mathbf{V}) = \|\mathbf{X} - \mathbf{AVB}'\|_F^2, \tag{3}$$

is minimized, with $\|\mathbf{X} - \mathbf{AVB}'\|_F^2$ denoting the Frobenius norm of a matrix. To this end, two alternating least squares algorithms (ALS) have been proposed: PENCLUS (Both and Gaul, 1985, 1987; Gaul and Schader, 1996), and the Clusterwise ALS approach of Baier et al. (1997). The main difference between these two approaches is that PENCLUS works with a penalized (auxiliary) loss function and operates in a (relaxed) continuous space, whereas Clusterwise ALS optimizes the original biclustering loss function in Eq. (3) directly and operates in the original solution space. Wilderjans et al. (2013b) evaluated the performance of these ALS algorithms in a simulation study.

The simulation study involved a full factorial design that was set up as follows. Data sets $\mathbf{X}$ were generated by independently generating true object cluster membership matrices $\mathbf{A}^{true}$, true variable cluster membership matrices $\mathbf{B}^{true}$, true weight matrices $\mathbf{V}^{true}$, and noise $\mathbf{E}$. We refer to Wilderjans et al. (2013b) for details on how the entries of $\mathbf{A}^{true}$, $\mathbf{B}^{true}$, $\mathbf{V}^{true}$, and $\mathbf{E}$ were drawn. A data set $\mathbf{X}$ was then obtained as $\mathbf{X} = \mathbf{T} + \mathbf{E}$, with $\mathbf{T} = \mathbf{A}^{true}\mathbf{V}^{true}(\mathbf{B}^{true})'$ and $\mathbf{T}$ being the true underlying clustering model. In the data set generation, six factors were manipulated: the shape of $\mathbf{X}$ (three levels), the number of object clusters (three levels), the number of variable clusters (three levels), the amount of object cluster overlap (three levels), the amount of variable cluster overlap (three levels), and the amount of noise (five levels). Regarding the shape of $\mathbf{X}$, this was defined in terms of the ratio I/J of the number of objects to the number of variables, and took values of 1/4, 1, and 4. The total number of entries in $\mathbf{X}$ was kept fixed at 4096, which implied three different values for $I \times J$: 32×128, 64×64, 128×32. Regarding the number of object and variable clusters, these both took values of 2, 3, and 4 (which can be considered a rather restricted set of numbers of clusters). The amount of both object and variable cluster overlap was defined as the probability of an object (variable) belonging to more than one object (variable) cluster, and was put equal to 25%, 50%, or 75%. Regarding the amount of noise, this factor was defined as the proportion ϵ of the total variance in the data $\mathbf{X}$ accounted for by $\mathbf{E}$ and was either 0, 0.15, 0.30, 0.45 or 0.60. Note that the shape of $\mathbf{X}$ is a characteristic of type (b) (known in data-analytic practice, see Table 1), whereas the other factors manipulated in the simulation study are of type (a) (i.e., not known in data-analytic practice).

All design factors were fully crossed, which yielded 1215 combinations. For each combination, 10 replicates were generated, yielding $N = 12\,150$ simulated data sets in total. Subsequently, each simulated data set $\mathbf{X}$ was analyzed with PENCLUS and Clusterwise ALS, resulting in estimates $\hat{\mathbf{A}}$, $\hat{\mathbf{B}}$, $\hat{\mathbf{V}}$ for both algorithms. Outcome variables in the simulation study included a measure of recovery of $\mathbf{T}$ by $\mathbf{M} = \hat{\mathbf{A}}\hat{\mathbf{V}}\hat{\mathbf{B}}'$, recovery of the true overlapping clustering $\mathbf{A}^{true}$ by $\hat{\mathbf{A}}$ (object cluster

recovery), recovery of the true overlapping clustering $\mathbf{B}^{true}$ by $\hat{\mathbf{B}}$ (variable cluster recovery), and recovery of the weights matrix $\mathbf{V}^{true}$ by $\hat{\mathbf{V}}$. In order to not confound the evaluation of the outcome variables with a possibly inadequate model selection procedure for selecting the number of object and variable clusters, in the simulation study the algorithms were assumed to be applied with a perfect model selection procedure. As such, the algorithms were applied to the simulated data sets with the number of object and variable clusters put equal to their true values; hence, the recovery of the true structure underlying the data sets could be evaluated in a meaningful way. In the present application, we focus on goodness-of-recovery of the weights matrix $\mathbf{V}^{true}$ (GOW) as outcome variable. To evaluate GOW, the following measure was calculated:

$$1 - \frac{\sum_{k=1}^{K}\sum_{l=1}^{L}(\hat{v}_{kl} - v_{kl}^{true})^2}{\sum_{k=1}^{K}\sum_{l=1}^{L}(v_{kl}^{true} - \bar{v}^{true})^2}, \tag{4}$$

with $\bar{v}^{true}$ the average value of $\mathbf{V}^{true}$. GOW was then defined as the minimum value of Eq. (4) over all possible row and column permutations of $\hat{\mathbf{V}}$; this measure takes values in the interval $(-\infty, 1]$, with a value of 1 meaning perfect recovery.

3.2. Analysis strategy

We estimated optimal data-analytic regimes through the stepwise procedure proposed in Section 2.4. That is, as a first step, we estimated the optimal data-analytic regime using all manipulated characteristics. As a second step, we subsequently estimated proxies for the characteristics that appeared to be relevant in the estimated regime resulting from Step 1 but that would be unknown to practitioners outside a Monte Carlo context. That is, we replaced the relevant characteristics in cell (a) of Table 1 by related variables in cell (d), and we subsequently estimated an optimal data-analytic regime involving only characteristics that are known in data-analytic practice (cells (b) and (d)).

In order to estimate optimal data-analytic regimes using all manipulated characteristics, we first constructed a contrast function $C(\mathbf{X}_i)$ for $i = 1, \ldots, N$. As contrast function $C(\mathbf{X})$, we calculated for each data set the difference in GOW between Clusterwise ALS and PENCLUS. Once we obtained the contrast function for each data set, $C(\mathbf{X}_i)$, we derived the class labels $Z_i = I[C(\mathbf{X}_i) > 0]$, and the weights $W_i = |C(\mathbf{X}_i)|$ for each data set, to obtain the classification data set $\{Z_i, \mathbf{X}_i, W_i\}$.

Subsequently, we estimated $g^{\text{opt}}(\mathbf{X}_i)$ as $\hat{g}^{\text{opt}}(\mathbf{X}_i) = \arg\min_{g\in\mathcal{G}} \sum_{i=1}^{N} W_i[Z_i - g(\mathbf{X}_i)]^2$, where the minimization is across all regimes in the class of regimes under study. For this minimization, we used classification trees (e.g., Hastie et al., 2009), which yield tree-based data-analytic regimes; such regimes are of particular interest as they go with a straightforward and most insightful representation of the decision structure underlying the regimes in question. In particular, we focused on regimes based on a tree structure with binary splits defined in terms of dichotomized characteristics of the data and the data-generating mechanism. The splits result in a finite set of nodes, which comprises a root node, internal nodes and leaves (or terminal nodes). Data-analytic regimes that rely on such a tree structure imply that decisions with regard to assignment of the methods under study are based on leaf membership. Let $S_\ell (\ell = 1, \ldots, L)$ be the leaves associated with such a tree, with the leaves being defined in terms of conjunctions of thresholded characteristics of the data and the data-generating mechanism, such as, for example, $[(X_1 > b)$ and $(X_2 \leq c)]$. A tree-based data-analytic regime g_T can then be formalized as a function, $g_T : \{S_1, \ldots, S_\ell, \ldots, S_L\} \rightarrow \{0, 1\}$.

The procedure for estimating tree-based data-analytic regimes was implemented in R, using the `rpart` package for fitting a classification tree (Therneau et al., 2013). We input the classification data set $\{Z_i, \mathbf{X}_i, W_i\}$ into `rpart` and adopted the default settings, except that we put the weights equal to the entries of W_i. The resulting classification tree was finally pruned back using the cross-validation procedure described by Hothorn and Everitt (2009).

Finally, if the estimation would end up in a nontrivial data-analytic regime, this would suggest that the data sets are involved in a qualitative or disordinal interaction, implying that for some types of data PENCLUS would outperform Clusterwise ALS, whereas for other types it would be the other way around. In this case, we will test the hypothesis that the interaction is disordinal, indeed.

3.3. Results and discussion

The classification tree resulting from the first step is displayed in Fig. 1. It appears that for data sets with a higher amount of noise one should preferably use Clusterwise ALS. This also holds for data sets with a lower amount of noise if the true number of object clusters is small. On the other hand, for data sets with a lower amount of noise one should preferably use PENCLUS if the true number of object clusters is not too small. We may thus conclude that the most important moderators of the differential performance of Clusterwise ALS and PENCLUS are amount of noise and number of object clusters. Regarding the number of object clusters, this result may be related to the fact that PENCLUS operates in a (relaxed) continuous space while Clusterwise ALS operates in the original solution space. When there is only a small number of underlying clusters, that is to say, when the original solution space is small, operating in this space appears to be better than operating in some (relaxed) continuous space. However, when there are many underlying clusters (i.e., when the original solution space is large), Clusterwise ALS may not be moving in an efficient way throughout that space and may get lost in it. Making use of

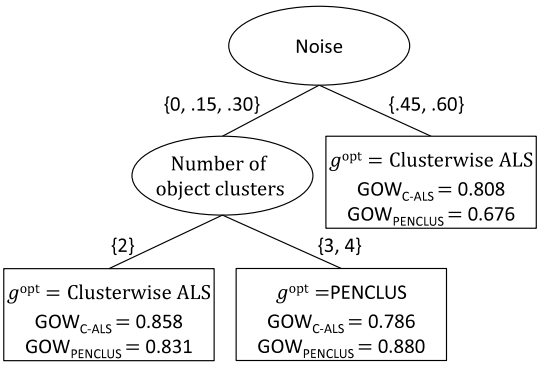


Fig. 1. Optimal data-analytic regime for the additive biclustering data resulting from the newly proposed method using classification trees. The ellipses in the figure represent the internal nodes containing the split variables, with the corresponding split point shown below each ellipse. The upper ellipsis represents the root node, which corresponds to the complete set of data sets from the benchmarking study. The rectangles represent the leaves of the tree, with each rectangle containing the assigned method, and the conditional mean outcomes under Clusterwise ALS (C-ALS) and PENCLUS.

the test for disordinal interactions developed by Gail and Simon (1985), we can reject the null hypothesis of no disordinal interaction at significance level $\alpha = 0.001$.

As the true number of object clusters and the amount of noise are manipulated characteristics of the data that are not known in data-analytic practice (cell (a) of Table 1), we identified proxies for these characteristics in order to estimate a data-analytic regime that can be usable in data-analytic practice. (For clarity's sake, these proxies are characteristics in cell (d) of Table 1.) To identify a proxy for the number of object clusters we fitted a one-mode clustering model (viz., an additive profile clustering model, ADPROCLUS: Depril et al., 2008; Wilderjans et al., 2011) to the simulated data sets along with an automated convex hull-based procedure (Ceulemans and Kiers, 2006; Wilderjans et al., 2013a) to determine the number of underlying object clusters. The latter number then serves as proxy for the number of object clusters K^{proxy}. It should be noted that this proxy was used in the estimation of the second data-analytic regime but not for the number of object clusters used in the analyses of the simulated data sets with Clusterwise ALS and PENCLUS. This means that the proxy was only used to decide on the optimal type of method, and not to choose the number of clusters in the application of this method. It should further be noted that, in practice, a model selection procedure should be specified to estimate the number of object and variable clusters in order to analyze a data set with Clusterwise ALS or PENCLUS.

To identify a proxy for the percentage of noise in the data, we relied on the object cluster membership matrix $\hat{\mathbf{A}}$ that resulted from the ADPROCLUS analysis mentioned above with K^{proxy} object clusters. To obtain also a variable cluster matrix $\hat{\mathbf{B}}$ for each data set, we subjected the transposed data $\mathbf{X}^T$ to ADPROCLUS, while again using the automated *CHull* procedure to select the number of variable clusters. We subsequently calculated $\hat{\mathbf{V}}$ as

$$\hat{\mathbf{V}} = (\hat{\mathbf{A}}^T\hat{\mathbf{A}})^+\hat{\mathbf{A}}^T\mathbf{X}\hat{\mathbf{B}}(\hat{\mathbf{B}}^T\hat{\mathbf{B}})^+,$$

with $(\mathbf{P})^+$ being the Moore–Penrose pseudoinverse of P. Next, we calculated as a proxy for the percentage of noise in the data,

$$\varepsilon^{\text{proxy}} = \frac{\|\mathbf{X} - \hat{\mathbf{A}}\hat{\mathbf{V}}\hat{\mathbf{B}}^T\|^2}{\|\mathbf{X}\|^2},$$

where $\|\mathbf{P}\|^2$ indicates the squared matrix norm of $\mathbf{P}$.

The classification tree resulting from the second analysis with the proxies for 'amount of noise' and 'true number of object clusters' as predictors is displayed in Fig. 2. Interestingly, this tree has the same structure as the tree based on the originally manipulated characteristics, which gives confidence in the result and the quality of the estimated proxies. The expected outcome of the estimated optimal data-analytic regime in Fig. 2 using Eq. (1) is 0.844. For comparison, the expected outcome would equal 0.809 if the marginally best method, Clusterwise ALS, were applied to all 12 150 data sets, and 0.788 if PENCLUS were applied to all of them. This means that, although the strategy of analyzing all data sets with Clusterwise ALS would result in an improved outcome relative to analyzing them with PENCLUS, there is added benefit to analyze them on the basis of the estimated optimal data-analytic regime.

By way of comparison, we also estimated optimal data-analytic regimes on the remaining three outcome variables included in the simulation study by Wilderjans et al. (2013b) (the trees are given in the Appendix). It appears that the estimated regime based on outcome variable GOW (Fig. 1) and the estimated regime based on the recovery of $\mathbf{T}$ by $\mathbf{M} = \hat{\mathbf{A}}\hat{\mathbf{V}}\hat{\mathbf{B}}'$ imply two (almost) identical splits on 'amount of noise' and 'true number of object clusters'. The two estimated regimes based on object cluster recovery and variable cluster recovery are also interrelated in many ways. However, these regimes are considerably more complex compared to the regimes based on GOW and on the recovery of $\mathbf{T}$. This nicely illustrates that the optimal regime may depend on the chosen outcome measure.

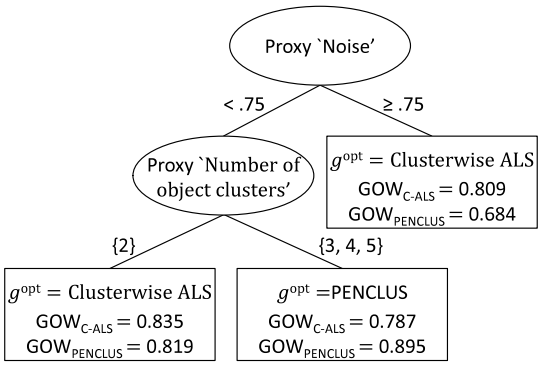


Fig. 2. Optimal data-analytic regime with replacement of noise and true amount of object clusters by proxies that are known in data-analytic practice.

A final remark pertains to the classification tree method that we used to estimate the data-analytic regimes in the application. Such methods are cursed by a limitation: Their search space is very large by the large number of possible split variables and the large number of possible split points per variable. This may result in instable solutions. There are several ways in which the instability of classification trees can be quantified. Examples include using a bagging procedure, where a set of trees is grown based on different bootstrap samples or subsamples of the data (Breiman, 1996), and subsequently calculating a measure of instability on this set of trees (Chipman et al., 1998). A previously suggested solution for the instability problem is the use of random forests. However, a key advantage of tree-based regimes (i.e., an insightful representation of the underlying decision structure) then gets lost. An interesting potential course of action to overcome this could be to look for a precis of a random forest in terms of a single summary tree.

4. General discussion

In the present paper we introduced a novel conceptual framework and methodology, inspired by principles of optimal treatment regime estimation, to derive optimal data-analytic regimes from benchmarking studies with simulated data. Over and above the search of a universally winning method, this methodology provides the user with a more refined answer to questions of benchmarking in terms of how the winning method may vary over patterns of characteristics of the data and the data-generating mechanism.

In general, optimal regime estimation is fairly challenging due to the high risk of inferential errors. The underlying reason for this high risk is that (nontrivial) optimal regimes critically depend on qualitative or disordinal interactions, which imply that for some types of data, method A outperforms B, whereas for other types it is the other way around. This type of interactions is typically hard to replicate (Rothwell, 2005). The reason for this does not reside in particular methods, but in the mere fact that a reliable estimation of qualitative or disordinal interactions requires (very) large sample sizes (Lee et al., 2015). A particular advantage of benchmarking studies with simulated data that may provide the basis for optimal data-analytic regime estimation is that they rather easily allow for large sample sizes, with the primary cost being computational in nature only (which in the present era of supercomputers has become less prohibitive than ever before).

A related issue is that exploratory post-hoc methods like the optimal regime estimation methods discussed in the present paper should be looked at as hypothesis-generating devices, the results of which are to be cross-validated in replication studies. Once again such replication studies are cheaper and therefore easier to realize in the benchmarking area compared to the area of randomized clinical trials. Two special types of replications deserve to be singled out at this point: (a) replication studies with stratified sampling schemes, with strata corresponding to the data types identified during the optimal data-analytic regime estimation, and (b) replication studies in which variables that were not experimentally manipulated in the original benchmarking study (cells (c) and especially (d) of Table 1) and that appeared to be involved in the regime resulting from that study, are experimentally manipulated in the follow-up replication study.

In this paper, we focused on benchmarking studies in which two methods are compared. We showed that in such cases, optimal data-analytic regimes can be estimated by using existing classification techniques that minimize some misclassification cost. In our case, the misclassification cost is data set-dependent and denotes the loss in outcome would a data set be assigned to its non-optimal method alternative. In the illustrative application, we further focused on methods to estimate optimal tree-based data-analytic regimes. Typical tree-based classification techniques assume misclassification costs to be the same for all objects that are to be classified (viz., in our case, for all data sets). As a solution to work around the issue of data set-dependent costs, we made use of the fact that in quite a few tree algorithms every object can be assigned a weight. By putting these weights equal to the individual misclassification costs, application of the tree-based classification techniques in question will imply a minimization of the total misclassification cost.

One may wonder how benchmarking studies in which more than two methods are compared could be dealt with, and whether our proposed approach to estimate data-analytic regimes could be extended to such cases. At this point it is relevant to know that (Doove et al., 2016) recently showed that optimal treatment regime estimation in the context of more than

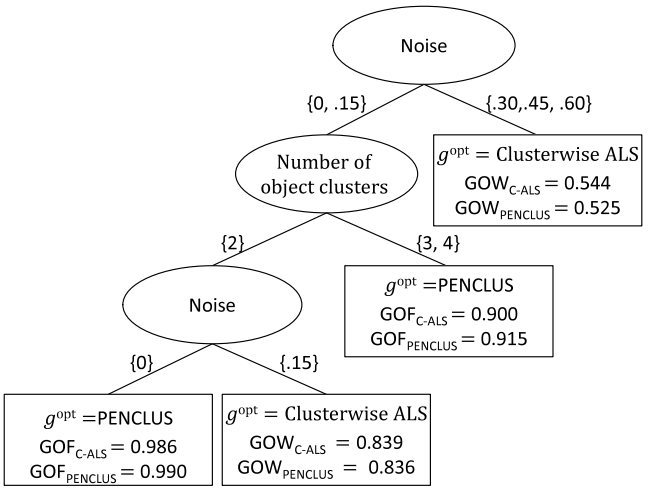


Fig. A.1. Optimal data-analytic regime for the additive biclustering data resulting from the newly proposed method using classification trees with as outcome measure a measure of reconstruction of **T** (GOF). The rectangles represent the leaves of the tree, with each rectangle containing the assigned method, and the conditional mean outcomes under Clusterwise ALS (C-ALS) and PENCLUS.

two treatment alternatives is also a classification problem in which some total misclassification cost is to be minimized. As in the binary case, the misclassification cost is object-dependent. In addition, the costs now also depend on the alternative to which an object is assigned. Importantly, to solve the resulting object- and alternative-dependent classification problem within a context of the estimation of tree-based regimes, it is no longer possible to rely on an object-based weighting. However, Doove et al. (2016) proposed an extension of a set of tree-based classification techniques to deal with object- and alternative-dependent classification costs, and implemented for this extension an R code add-on to a well-established package for classification tree estimation, `rpart` (the code of which can be obtained from the first author).

The proposed method may be further extended in two ways. First, as method acts as a within-subject variable in benchmarking studies, a difference score can be used as a straightforward estimator of the contrast function C. In optimal treatment regime estimation, however, the use of more involved estimators, which include an augmentation with a term borrowed from an outcome model, may be necessary (Zhang et al., 2012). It would be meaningful to explore whether a similar augmentation could still further improve the quality of estimated optimal data-analytic regimes.

Secondly, we discussed the estimation of optimal data-analytic regimes on the basis of data from benchmarking studies relying on "synthetic" Monte Carlo simulations. There are no principled objections against the application of the methodology proposed in the present paper to benchmarking studies based on empirical data (which could be considered a sample from a population of data sets: Boulesteix et al., 2015). Yet, two considerable pragmatic bottlenecks need to be taken into account: (a) The sample size of benchmarking studies with empirical data is typically prohibitively low, and an increase of that sample size is considerably more expensive than in the case of simulated data; (b) inferences may be hampered by correlations between data characteristics across the empirical data sets. Having taken these bottlenecks into account, a full estimation of optimal data-analytic regimes on the basis of benchmarking studies with empirical data does not look like a realistic option. As a more realistic alternative, the data sets of such studies could be used as a test bench to investigate optimal data-analytic regimes that were previously derived from benchmarking studies with simulated data sets. In case the application to the test bench would lead to unexpected discrepancies with the results of the simulated data sets, this could subsequently give rise to the setting up of new benchmarking studies with simulated data sets that mimic in one way or another the empirical data sets of the test bench (for an illustration of this principle, see Schepers et al., 2006). Otherwise, in case the primary interest of the researcher resides in outcome measures of a recovery type (as in our illustrative application), a straightforward test on empirical data is not possible either. Yet, in such cases, too, one could consider to set up "realistic simulation studies" with a test bench of simulated data sets that mimic in some way empirical data sets of interest.

A final remark pertains to the emphasis that has been put in this paper on the perspective of the practitioner who appreciates pragmatic rules to choose the best data-analytic method for a data set at hand given known characteristics of that data set (which implies quite some emphasis on the "what" of the optimal data-analytic regimes). Yet, the "why" of the regimes in question (i.e., the theoretical justification of why a particular method outperforms another one for certain types of data sets) should not be neglected at all. Insight into this "why" may simply lead to more sound regimes that can be extended more easily to new types of data sets.

Acknowledgments

The research reported in this paper was supported in part by the Research Fund of KU Leuven (GOA/15/003), and by the Interuniversity Attraction Poles programme financed by the Belgian government (IAP/P7/06).

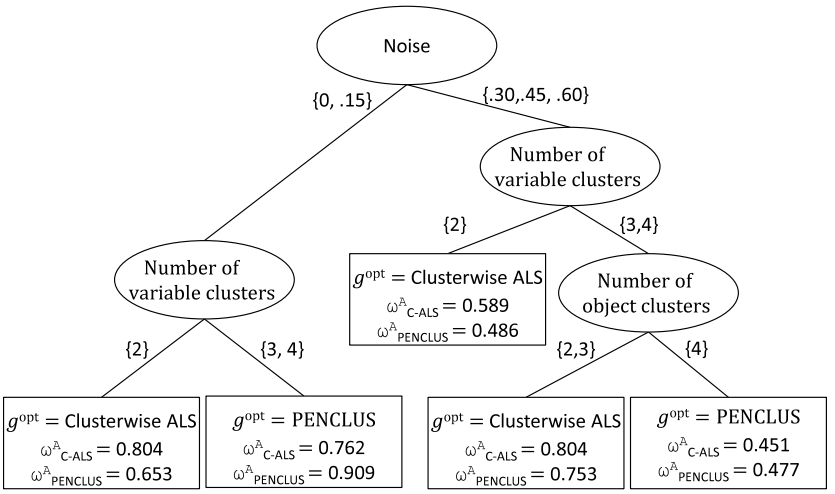


Fig. A.2. Optimal data-analytic regime for the additive biclustering data resulting from the newly proposed method using classification trees with as outcome measure object cluster recovery (ω^A). The rectangles represent the leaves of the tree, with each rectangle containing the assigned method, and the conditional mean outcomes under Clusterwise ALS (C-ALS) and PENCLUS.

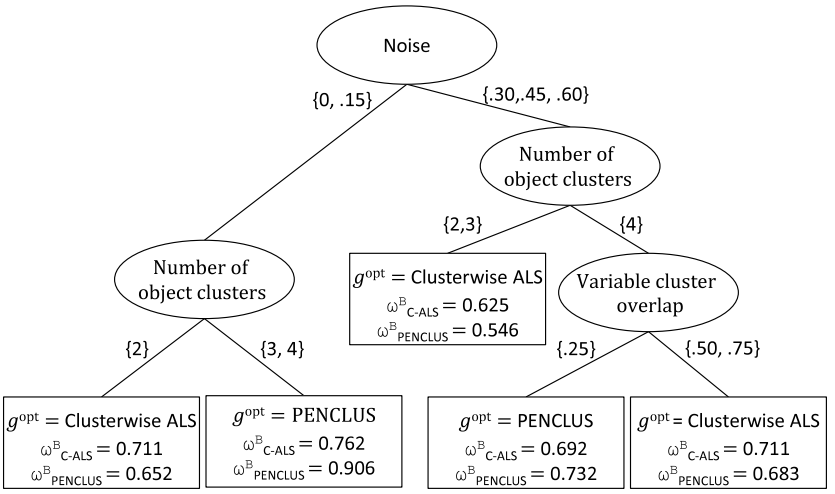


Fig. A.3. Optimal data-analytic regime for the additive biclustering data resulting from the newly proposed method using classification trees with as outcome measure variable cluster recovery (ω^B). The rectangles represent the leaves of the tree, with each rectangle containing the assigned method, and the conditional mean outcomes under Clusterwise ALS (C-ALS) and PENCLUS.

Appendix

In this Appendix, we present estimated optimal tree-based data-analytic regimes with as outcome measures a measure of reconstruction of $\mathbf{T}$ by $\mathbf{M} = \hat{\mathbf{A}}\hat{\mathbf{V}}\hat{\mathbf{B}}'$ (Fig. A.1), object cluster recovery (Fig. A.2), and variable cluster recovery (Fig. A.3).

References

Baier, D., Gaul, W., Schader, M., 1997. Two-mode overlapping clustering with applications to simultaneous benefit segmentation and market structuring. In: Klar, R., Opitz, K. (Eds.), Classification and Knowledge Organization. Springer, Berlin, Germany, pp. 557–566.

Both, M., Gaul, W., 1987. Ein vergleich zweimodaler clusteranalyseverfahren. Meth. Oper. Res. 57, 593–605.

Both, M., Gaul, W., 1985. PENCLUS: Penalty clustering for marketing applications. Discussion Paper No. 82, Institution of Decision Theory and Operations Research, University of Karlsruhe.

Boulesteix, A.-L., Hable, R., Lauer, S., Eugster, M.J.A., 2015. A statistical framework for hypothesis testing in real data comparison studies. Amer. Statist. 69, 201–212.

Breiman, L., 1996. Bagging predictors. Mach. Learn. 24, 123–140.

Brookes, S.T., Whitley, E., Peters, T.J., Mulheran, P.A., Egger, M., Davey Smith, G., 2001. Subgroup analyses in randomised controlled trials: quantifying the risks of false-positives and false-negatives. Health Technol. Assess 5, 1–56.

Ceulemans, E., Kiers, H.A.L., 2006. Selecting among three-mode principle component models of different types and complexities: A numerical convex hull based method. Br. J. Math. Stat. Psychol. 59, 133–150.

Chipman, H.A., George, E.I., McCulloh, R.E., 1998. Making sense of a forest of trees. In: Weisberg, S. (Ed.), 30th Symposium on the Interface, Fairfax Station. VA: Interface Foundation of North-America, pp. 84–92.
Depril, D., Van Mechelen, I., Mirkin, B.G., 2008. Algorithms for additive clustering of rectangular data tables. Comput. Statist. Data Anal. 52, 4923–4938.
DeSarbo, W.S., 1982. GENNCLUS: New models for general nonhierarchical clustering analysis. Psychometrika 47, 449–475.
Doove, L.L., Dusseldorp, E., Van Mechelen, I., 2016. Extension of a generic approach for estimating optimal treatment regimes to multivalued treatments, Manuscript submitted for publication.
Gail, M., Simon, R., 1985. Testing for qualitative interactions between treatment effects and patient subsets. Biometrics 41, 361–372.
Gaul, W., Schader, M., 1996. A new algorithm for two-mode clustering. In: Bock, H.-H., Polasek, W. (Eds.), Data Analysis and Information Systems: Statistical and Computational Approaches. Springer, Berlin, Germany, pp. 15–23.
Hastie, T., Tibshirani, R., Friedman, J., 2009. The elements of statistical learning: Data mining, inference, and prediction (2nd ed.). Springer, New York.
Hothorn, T., Everitt, B.S., 2009. A Handbook of Statistical Analyses Using R, second ed. CRC Press, Boca Raton.
Lagakos, S.W., 2006. The challenge of subgroup analyses-reporting without distorting. New Engl. J. Med. 354, 1667–1669.
Lee, S., Lei, M.-K., Brody, G.H., 2015. Confidence intervals for distinguishing ordinal and disordinal interactions in multiple regression. Psychol. Methods 20, 245–258.
Rothwell, P.M., 2005. Subgroup analysis in randomized controlled trials: Importance, indications, and interpretation. Lancet 365, 176–186.
Schepers, J., Van Mechelen, I., Ceulemans, E., 2006. Three-mode partitioning. Comput. Statist. Data Anal. 51, 1623–1642.
Shaffer, J., 1991. Probability of directional errors with disordinal (qualitative) interaction. Psychometrika 56, 29–38.
Therneau, T., Atkinson, B., Ripley, B., 2013. rpart: Recursive partitioning. R package version 4.1-3. http://CRAN.R-project.org/package=rpart.
Wilderjans, T.F., Ceulemans, E., Meers, K., 2013a. CHull: A generic convex hull based model selection method. Behav. Res. Methods 45, 1–15.
Wilderjans, T.F., Ceulemans, E., Van Mechelen, I., Depril, D., 2011. ADPROCLUS: a graphical user interface for fitting additive profile clustering models to object by variable data matrices. Behav. Res. Methods 43, 56–65.
Wilderjans, T.F., Depril, D., Van Mechelen, I., 2013b. Additive biclustering: A comparison of one new and two existing ALS algorithms. J. Classification 30, 56–74.
Zhang, B., Tsiatis, A.A., Davidian, M., Zhang, M., Laber, E., 2012. Estimating optimal treatment regimes from a classification perspective. Stat 1, 103–114.