Nonparametric estimation of customer segments from censored sales panel data

Johannes F. Jörg[1] · Catherine Cleophas[2]

Abstract

Specifically addressing different customer segments via revenue management or customer relationship management, lets firms optimize their market response. Identifying such segments requires analysing large amounts of transactional data. We present a nonparametric approach to estimate the number of customer segments from censored panel data. We evaluate several model selection criteria and imputation methods to compensate for censored observations under different demand scenarios. We measure estimation performance in a controlled environment via simulated data samples, benchmark it to common clustering indices and imputation methods, and analyse an empirical data sample to validate practical applicability.

Keywords Nonparametric statistics · Demand estimation · Demand segmentation · Censored panel data

Introduction

Divulging the demand structure from historical sales data is inherently difficult: On the one hand, a multitude of external factors, such as competing offers and seasonality, affect demand. On the other hand, inventory controls and pricing censor observed sales, so that sales do not reflect actual demand. This particularly challenges demand forecasting for revenue management and pricing, as reviewed by Azadeh (2013).

Nevertheless, accurately identifying customer segments provides a crucial foundation for decision-making. It enables firms to optimize their market response by targeting specific segments based on data rather than intuition.

In one business case, revenue management aims to optimize product offer sets and prices over a fixed sales horizon. For air travel, a tourism customer segment may book early in the sales horizon, exhibiting low willingness to pay and high flexibility with regard to departure times. In contrast, a business customer segment may book later, exhibiting high willingness to pay and a strong preference for a small range of departure times. By catering to different segments, revenue optimisation can determine the best number of reduced-fare tickets to offer to the tourism segment and the best number of high-fare tickets to reserve for the business segment (compare Talluri and Van Ryzin (2004), Sect. 11.1). In one example of related research, Meissner et al. (2013) extend a segment-based deterministic concave programme to maximize expected revenues for overlapping customer segments. The authors define segments via a consideration set of products and assume the number of segments and their structure to be given.

In another business case, customer relationship management aims to identify potential 'prospects', potential customers that are likely to react positively when being contacted in a particular way (see Linoff and Berry (2011), Chapter 4). Here, customer segmentation aims to differentiate, for instance, a segment of social-media affine customers, who are best contacted via social networks, versus a segment of more traditional-minded customers, who react best to postal mail and phone calls.

In a third example, Müller and Haase (2014) present a location planning model for retail facilities based on customer segmentation. The authors show that considering customer choices within segments improves the planning outcomes. They model customer choice with a multinomial logit model. The numerical study confirms improvements

✉ Johannes F. Jörg
johannes.ferdinand.joerg@rwth-aachen.de

Catherine Cleophas
cleophas@bwl.uni-kiel.de

1 Advanced Analytics, RWTH Aachen University
   Kackertstraße 7, 52072 Aachen, Germany

2 Service Analytics, Christian-Albrechts-Universität zu Kiel,
   Westring 425, 24118 Kiel, Germany

when considering customer characteristics per segment rather than averaging customer characteristics.

Customer segmentation seeks to answer four questions: 1. How many segments are there? 2. What is the market share per segment with regard to the general population? 3. What is the characteristic behaviour per segment? 4. How can an individual be identified as belonging to a segment? In this paper, we concentrate on the first question. The proposed algorithm contributes a way to answer this question based on sales observations, supporting further work on all further questions that assume a given number of segments.

Existing approaches to demand estimation neglect the challenge of estimating the number of segments without prior assumptions. For example, Talluri and Van Ryzin (2004) state that, whilst customer segmentation is "well suited to analytical methods", it is "still based primarily on managerial judgement and intuition" (p. 580).

Assuming the wrong number of customer segments may lead to various problems: Underestimation may result in aggregating too much information, such that characteristics of specific segments are lost. This, in turn, may cause lost business opportunities, since any further decision-making cannot fully address those characteristics. Overestimation may lead to overfitting. As an extreme example, when assuming that each product is bought exclusively by one specific customer segment (independent demand), customers' substitution behaviour will not be captured accurately.

Furthermore, current application-oriented approaches to estimate demand are predominantly *parametric*. Examples include naive heuristics (Weatherford and Pölt 2002; Saleh 1997), expectation maximisation (Little and Rubin 2014; Salch 1997), or projection detruncation (Skwarek 1996). However, parametric estimation requires experts to specify the underlying distribution. Misspecified distributions may inaccurately represent the data and therefore lead to estimation biases or efficiency loss (Härdle and Mammen 1993; Agresti et al. 2004; Litière et al. 2008).

Our contribution aims to support and supplement existing forecasting and optimization procedures that require knowledge about the number of customer segments. To this end, we present a *nonparametric approach* to estimate the number of customer segments from censored panel data. Specifically, we model customer segments via *finite mixture models*. These explain sales observations as compositions of several, possibly overlapping mixture components. They provide a general framework to implement any type of overall distribution that consists of several sub-distributions and are therefore fitting to estimate the number of customer segments. Thus, they can be used to model and approximate any arbitrary setting. An introduction to finite mixture models can be found in McLachlan and Peel (2000).

Given minimal assumptions and no external knowledge, the proposed approach employs sales panel data to abandon the need for assuming a specific underlying distribution. It requires only panel data of two consecutive buying decisions, which is essentially the most limited type of panel data. Thus, it can evaluate markets with scarce panel data, such as airline bookings. In addition, we consider situations when the firm did not offer a product, thereby censoring the sales data. The result provides the basis for further parametric or nonparametric techniques to estimate the customer segment characteristics.

The proposed approach only relies on one unique identifier, e.g., a customer ID, requiring no further customer information. In the light of ongoing discussions concerning customer privacy and data storage, using further customer specific information for revenue optimisation may meet legal restrictions as highlighted by the new European General Data Protection Regulation. This motivates methods that require comparatively little data, as exemplified here.

As our primary theoretical contribution, we extend a nonparametric identification of multivariate mixtures to be applicable to censored panel data. Our primary practical contribution is to evaluate a nonparametric estimation approach for censored sales panel data. We perform extensive computational experiments, confronting the proposed approach with different demand models and types of panel data.

This paper is structured as follows: section "Related literature and theoretical background" reviews existing research on nonparametric demand estimation, particularly in the context of revenue management. Section "Methodology" formalizes the proposed approach. Section "Results" evaluates the approach on simulated and empirical data sets. Finally, Section "Conclusion" discusses our findings and addresses managerial implications.

Related literature and theoretical background

This section reviews related research on customer segmentation and estimation, particularly in the context of revenue management. We also briefly discuss the more theoretical contributions as foundations of the newly proposed approach.

With regard to demand estimation for revenue management, (Queenan et al. 2007; Ratliff et al. 2008; Azadeh 2013) review uncensoring methods. Talluri (2009) discusses commonly used revenue management demand models and their limitations. The author proposes an approach which allows for a more flexible way to choose from demand distributions. Rusmevichientong et al. (2014) focus on improving expected revenue optimisation by ordering assortments according to the products of highest revenue. Kunnumkal (2014) proposes a sampling based dynamic programme with

randomization to minimize the assumptions on the customer choice model.

A promising nonparametric approach to model customer choice is presented in Farias et al. (2013). It defines segments by customers' preferences over a list of products, which are estimated from sales transaction data. The authors simulate data for different underlying models such as multinomial logit, nested logit, and mixed multinomial logit. Their approach produces new revenue estimates, which they compare to revenue estimates derived from a known underlying logit model. However, this nonparametric approach models the number of customer segments only implicitly.

Another significant recent contribution to nonparametric demand estimation is documented in Van Ryzin and Vulcano (2015). The proposed approach requires historical sales and availability data as well as an initial set of customer segments. Rather than employing common multinomial logit or nested logit models, it characterizes customer segments by preference lists over offered products. An expectation–maximisation procedure iteratively adds new customer segments. However, the authors still rely on assumptions about customer segments, e.g., when initially creating one segment per product. Haensel and Koole (2011) propose a similar approach, but do not address the creation of customer segments.

Jagabathula and Vulcano (2015) present a general framework for the nonparametric estimation of preferences from panel data. The model characterizes demand as a directed acyclic graph (DAG). One of the main assumptions is "inertia of choice", i.e., once a customer bought a product, they will be unlikely to consider other alternatives when next purchasing a product of this type. "Trigger events", such as the non-availability of a product or product promotions, cause customers to reconsider the initial choice. Clustering these DAGs creates a predetermined number of segments.

Azadeh et al. (2015) propose another nonparametric approach to estimate customer choice and the resulting revenue. They represent customer choice by a multinomial logit model, in which utilities are product-based standalone variables, to circumvent parameterising utilities. Customer segments are characterized by their consideration set and probabilities of purchase. The authors validate their method on simulated data sets including 24 different demand parameter scenarios. They find that their approach is computationally efficient whilst still maintaining acceptable approximations. An algorithmic comparison between existing solvers and the proposed approach can be found in Azadeh et al. (2015). However, whilst being applied to a setting similar to the one we consider, the approach does not compute the number of customer segments.

Our work complements further estimation and forecasting steps as presented in Müller and Haase (2014), Van Ryzin and Vulcano (2015), and Jagabathula and Vulcano (2015).

Rather than assuming a given number of segment, e.g., by expert intuition, we estimate this number directly from the data. The result can inform the initial set of customer segments or the number of DAG clusters. In addition, we benchmark several approaches to determine the best number of clusters based on the common k-means clustering algorithm. We also consider alternative data imputation approaches and suggest ways of choosing model selection criteria.

The estimation approach proposed here extends the work of Kasahara and Shimotsu (2014) and applies it to a practical demand setting. Kasahara and Shimotsu (2014) present a method to estimate a lower bound for the number of components, i.e., the number of customer segments, for a finite mixture model. The method relies on uncensored panel data reporting two consecutive choices from alternatives defined by a continuous attribute. It aggregates panel data into a probability matrix and gives a decomposition of it. The authors use the idea that the non-negative rank of the matrix is equal to the number of components (as shown in Cohen and Rothblum (1993)). To identify the rank, they apply sequential hypothesis testing based on a rank statistic and compare it to an Akaike and a Baysian information criterion for different random variable sets.

In this paper, we present and discuss the approach for choices from discrete alternatives rather than alternatives characterized by a common continuous attribute. To apply sequential hypothesis testing in our setting, we apply the rank statistic by Robin and Smith (2000), which is based on the characteristic roots of a sample matrix. Here, the idea is that the rank of a matrix is equal to the non-zero characteristic roots. We also propose imputation rules to make the estimation approach applicable to censored sales panel data. We compare estimation performance given sequential hypothesis testing, the Akaike, and the Baysian information criterion.

Methodology

The proposed approach combines rank estimation of panel data as introduced by Kasahara and Shimotsu (2014), a characteristic root statistic by Robin and Smith (2000), and imputation rules to incorporate censoring. We extend (Kasahara and Shimotsu 2014) to study choices from discrete products.

The section first introduces the basics of finite mixture models. Subsequently, we present the rank estimation approach for data where observed choices considered discrete attributes and explain the estimation procedure, including the rank statistic and model selection. Finally, we extend the approach to account for censored observations of discrete attributes by introducing possible data imputation rules.

Finite mixture models

Finite mixture models provide a flexible way of modelling arbitrary distributions. They assume that the overall population, which underlies the observations, consists of smaller sub-populations. This assumption frequently applies in the real world, e.g., when modelling height of female and male persons, purchase behaviour of customers in a retail store, or observations of different, overlapping light sources in deep space astronomy. Employing finite mixture models lets us estimate properties of sub-populations from observations of the whole population.

Consider an observation set of N random vectors $\mathbf{X} = (\mathbf{X^1}, \dots, \mathbf{X^N})$, where each T-dimensional vector $\mathbf{X^i}$ is composed of random variables $X_1^i, \dots, X_T^i$. Each vector can represent the sales data for customer i over all observed decisions T. Observations may consist of *discrete attributes*, such as the product bought, or *continuous attributes*, such as the price paid or the product's quality. In this setting, we focus on discrete attributes.

Let $p(\mathbf{x^i})$ be the probability mass function of the random vector $\mathbf{X^i}$. For M so-called components, $p(\mathbf{x^i})$ can be expressed as

$$p(\mathbf{x^i}) = \sum_{m=1}^{M} \pi_m p_m(\mathbf{x^i}). \tag{1}$$

In our context, M is the number of customer segments. Here, $p_m(\mathbf{x^i})$ are component masses and $\pi_m \in [0, 1]$ for $m \in \{1, \dots, M\}$ are mixture weights with

$$\sum_{m=1}^{M} \pi_m = 1.$$

This ensures that observations consist of exactly M components. Since $p_m(\mathbf{x^i})$ are densities, it follows that $p(\mathbf{x^i})$ also defines a density. It is called M-component finite mixture density. p_m are distributions over the attributes of observed sales depending on the customer segment m.

Note that the proposed estimation approach relies on the formulation given by (1). Therefore, it does not rely on any distributional assumptions, as the densities p_m are arbitrarily chosen. The formula can easily be used to model popular choice models, such as the multinomial logit models, by choosing the densities $p_m(\mathbf{x_i})$ accordingly, but is not restricted to such models.

Estimation from observations with discrete attributes

Kasahara and Shimotsu (2014) show that using their rank estimation procedure on panel data yields a lower bound for the number of segments. However, the authors focus mainly on observed choices based on continuous attributes. To estimate the number of customer segments based on choices between discrete products, we adjust the approach as presented in the following.

We consider sales panel data including two consecutive observations $x_t^i, t \in \{1, 2\}$, per N uniquely identified customers i. Observations x_t^i take values in a finite discrete set S, which represents all offered products, and S_1 and S_2 represent the offer set of products for observations $t \in \{1, 2\}$.

In our context, segments represent groups of individual customers, with similar buying behaviours. We model this by a consideration set of the form $\{s_1, \dots, s_k\} \subset S_1 \cup S_2$ including k products. All customers in one segment exclusively buy products from this consideration set.

By definition, panel data refers to repeated observations of the same customer's purchase decisions. For instance, the IRI Academic Data Set (Bronnenberg et al. 2008) contains panel data of different grocery categories over the course of several years. For this data set, Jagabathula and Vulcano (2015) report that in almost all considered categories, the ratio of customers with more than a single transaction exceeds 50%. In these categories, the average transactions of customers with more than one transaction is higher than 3.07 and may even be as high as 19.81.

To model repeated customer choices, we suggest a finite mixture model following the form of (1). We assume that both choice probabilities $p_{m,1}$ and $p_{m,2}$ are independent and thus factorize,

$$P(x_1, x_2) = \sum_{m=1}^{M} \pi_m p_{m,1}(x_1) p_{m,2}(x_2), \tag{2}$$

such that $\sum_m \pi_m = 1$. $P(x_1, x_2)$ denotes the probability to observe a sample $(x_1, x_2) \in S_1 \times S_2$.

The model is composed of a mixture of M segments with mixture weights π_m and a product of time-dependent and segment-dependent masses $p_{m,t}(x_t)$. In the computational study, we assume that $p_{m,1}$ and $p_{m,2}$ are constant over time. Nevertheless, the model can also account for customer behaviour that evolves over time. Causes for this may include new market innovations, economic cycles, or a shift in market structures.

Example Let customers from two customer segments m_1 and m_2 make two consecutive choices from ten products $S = \{1, 2, 3, \dots, 10\}$. At the time of the first choice, the firm offers products 1 to 6; at the time of the second choice, it offers products 2 to 9. Customers from segment m_1 choose products from the consideration set $\{1, \dots, 7\}$ with uniform probability, whereas customers from segment m_2 consider products $\{8, 9, 10\}$. The two customer segments have the following market shares: $\pi_1 = 0.75$ and $\pi_2 = 0.25$. Table 1 illustrates this example of panel data of eight customers.

Table 1 Example of panel data describing two consecutive observations of eight customers choosing from ten offered products

Customer i	1	2	3	4	5	6	7	8
Observation x_1^i	6	4	1	4	8	5	10	1
Observation x_2^i	2	6	2	3	9	5	9	3
Customer segment	m_1	m_1	m_1	m_1	m_2	m_1	m_2	m_1

For instance, $x_2^4 = 3$ describes that customer 4, belonging to segment m_1, bought product 3 at the second observation. Note that the empirical data does not assign customers to segments – this is the task of the further estimation steps once the number of segments has been computed.

When customer x^1, from segment m_1, purchases first product 6 and subsequently product 2, the resulting panel data is $(x_1^1, x_2^1) = (6, 2)$. The finite mixture model represents this as a single event ($x_1 = 6$, $x_2 = 2$). The complete probability of observing event ($x_1 = 6$, $x_2 = 2$) is given by

$$f(x_1 = 6, x_2 = 2) = \sum_{m=1}^{2} \pi_m f_m(x_1 = 6, x_2 = 2)$$

$$= \pi_1 \cdot \frac{1}{7} \cdot \frac{1}{7} + \pi_2 \cdot 0 = \frac{1}{98} \approx 1\%$$

For simplicity, we assume without loss of generality that $S_1 = \{1, \ldots, p\}$, $S_2 = \{1, \ldots, q\}$, and $p \geq q$, which means that we can enumerate products for each consecutive choice. We characterize (2) by the following ($p \times q$)-matrix

$$P = \begin{pmatrix} P(X_1 = 1, X_2 = 1) & \cdots & P(X_1 = 1, X_2 = q) \\ \vdots & \ddots & \vdots \\ P(X_1 = p, X_2 = 1) & \cdots & P(X_1 = p, X_2 = q) \end{pmatrix}. \quad (3)$$

Let $\mathbb{1}_X(\cdot)$ denote the indicator function on X defined by

$$\mathbb{1}_X(x) = \begin{cases} 1, & \text{if } x \in X, \\ 0, & \text{else.} \end{cases}$$

From observations, we create an empirical version of matrix P

$$\hat{P}(x_1, x_2) = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}_{(x_1, x_2)}(x_1^i, x_2^i). \quad (4)$$

$\hat{P}(x_1, x_2)$ counts how often a specific observation (x_1^i, x_2^i) occurred and divides the result by the total number of observations N. This corresponds to an empirical probability matrix, in which each of the entries P_{x_i, x_j} indicates the empirical probability of an occurrence of an observation (x_i, x_j).

The estimation approach proposed here aims to assess the number of segments M. It assumes that this number is given by the lower bound derived from the rank of P. The approach estimates the rank of P using the rank statistic by

Robin and Smith ([2000](#)) on the empirical matrix $\hat{P}$. The rank of a given matrix is equal to the number of non-zero eigenvalues or characteristic roots. Thus, the so-called characteristic root test statistic (*CRT*) is based on the sorted eigenvalues $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \cdots \geq \hat{\lambda}_p$ of the Gram matrix $\hat{P}\hat{P}^T$, where $\hat{P}^T$ denotes the transpose of $\hat{P}$. Note that, for real matrices $\hat{P}$, the rank (*rk*) does not change when multiplying them with their transpose, i.e., $rk(\hat{P}) = rk(\hat{P}\hat{P}^T) = rk(\hat{P}^T\hat{P})$. Therefore, $\hat{\lambda}_{q+1} = \cdots = \hat{\lambda}_p = 0$ since $q \leq p$. *CRT* is defined as

$$CRT(m) = N \sum_{i=m+1}^{q} \hat{\lambda}_i. \quad (5)$$

If the real rank of matrix $\hat{P}$ is M, for $m < M$ follows $CRT(m) \to \infty$, as $N \to \infty$, since at least one eigenvalue in (5) is non-zero. For $m = M$, $CRT(m)$ has an asymptotic distribution given by a weighted sum of squared independent standard normal variables.

The rank of $\hat{P}$ has an upper bound of $\min\{|S_1|, |S_2|\}$. Therefore, this approach can identify at most as many segments as there are distinct products. Circumventing this limitation requires observing more purchase decisions per individual. For instance, for four observed decisions with product assortments S_1, S_2, S_3, and S_4, an equivalent to (3) results when grouping products of two decisions together, e.g., creating a ($|S_1| \cdot |S_2| \times |S_3| \cdot |S_4|$)-matrix.

The proposed approach merely estimates the number of customer segments in a market, it does not estimate their proportion or their choice behaviour. Therefore, it does not assign customers to segments or track shifts in the number of customers per segment. The approach only indicates a change in the number of segments when it is applied to a new set of observations and a whole customer segment dissipated over time without being replaced by a new segment.

Model selection

We consider three alternative ways of estimating the matrix rank M using the rank statistic *CRT*. First, we describe a sequential hypothesis testing (SHT). Subsequently, we adapt the Akaike (AIC) and Baysian information criteria (BIC) to the nonparametric setting.

From the empirical matrix $\hat{P}$, $\hat{F}_M^{CRT}$ can consistently estimate the asymptotic distribution function of $CRT(M)$. Here, $\hat{F}_M^{CRT}$ is the distribution function of $\sum_{i=1}^{(p-M)(q-M)} \hat{\gamma}_i Z_i^2$, where $\hat{\gamma}_i$ are eigenvalues of an auxiliary matrix (see the results of Robin and Smith (2000 in Appendix A.1) and Z_i are standard normally distributed random variables.

Based on this, we can formulate a hypothesis testing procedure. Let H_0 denote the null hypothesis, stating that the rank of matrix P is m. Let H_1 denote the alternative hypothesis, stating that the rank of P is larger than m. Therefore, with $H_0 : rk(P) = m$ versus $H_1 : rk(P) > m$, sequentially performing one-sided tests of H_0 against H_1 for $m = 0, 1, \ldots, q$, starting at $m = 0$, estimates rank M.

The first m for which H_0 is not rejected is the rank estimate $\hat{M}$:

$$\hat{M} = \min_{m \in \{0,\ldots,q\}} \{m : CRT(m) \geq \hat{q}_{1-\alpha_N}^i, i = 0, \ldots, m-1, \text{ and } CRT(m) < \hat{q}_{1-\alpha_N}^m \}.$$

Here, $\hat{q}_{1-\alpha}^m$ denotes the $1 - \alpha$-quantile of the cumulative distribution function $\hat{F}_m^{CRT}$. These quantiles define critical values, against which the test statistic is measured. If the statistic $CRT(m)$ lies within the region defined by the quantiles, m is considered as a candidate for the rank of the corresponding matrix. The minimum $\hat{M}$ therefore chooses the rank m for which the $CRT(m)$ statistic falls below the quantile $\hat{q}_{1-\alpha_N}^m$ for the first time. α hereby indicates the probability of rejecting the true hypothesis $H_0 : rk(P) = m$. Asymptotically, Robin and Smith (2000) have shown that $\hat{M}$ is a weakly consistent estimator for M, i.e., under fitting conditions for α_N, $\hat{M}$ converges to M. Note that the estimated rank is bounded by the smallest dimension, here q, of matrix $\hat{P}$.

Thus, we established a sequential hypothesis testing procedure that enables us to reject estimates of m that are too small. The proposed approach relies on the assumption that this lower bound yields the number of segments.

In the nonparametric setting, the original formulations of AIC and BIC are not applicable: They involve likelihood functions, but the underlying distributions are not known. Therefore, we adapt these criteria as follows:

$$AIC = CRT(m) - 2(p-m)(q-m)\bar{\gamma}(m),$$
$$BIC = CRT(m) - \log(N)(p-m)(q-m)\bar{\gamma}(m),$$

with $\quad \bar{\gamma}(m) = \dfrac{1}{(p-m)(q-m)} \sum_{i=1}^{(p-m)(q-m)} \hat{\gamma}_i.$

Here, $\bar{\gamma}(m)$ is the average of the characteristic roots $\hat{\gamma}_i$. The estimated rank $\hat{M}$ is then chosen via

$$\hat{M}_{AIC} = \min_{m \in \{0,\ldots,q\}} AIC(m), \qquad \hat{M}_{BIC} = \min_{m \in \{0,\ldots,q\}} BIC(m).$$

"Technical details" section in Appendix states a proposition concerning the asymptotic behaviour of the information criteria as introduced in Kasahara and Shimotsu (2014).

Imputing censored sales data

Given scarce capacity or revenue maximising inventory controls, a firm may not offer every product at all times. The resulting limited offer sets censor observed customer choices. Thus, sales observations may not reflect the customers' preferred choice. In the following, we assume that censoring causes missing data, where customers choose to not purchase any of the offered products.

Furthermore, we assume that the existence of missing data is reported as a non-purchase decision, following, e.g., assumptions established in Van Ryzin and Vulcano (2015). The resulting data distinguishes a period without customer arrival from a period with a customer arrival that did not lead to a purchase. In a practical setting, non-purchases can be observed, e.g., by analysing click-streams. We therefore assume that a customer visits a retailer with the intention of buying a product, not just to browse products. This assumption seems reasonable when considering frequently bought products, such as groceries. For products such as cars or airline tickets, click-stream data requires additional analysis to differentiate buying and browsing behaviour, as discussed for instance in Moe and Fader (2004). Furthermore, the proposed approach considers panel data, which can only be reported for identified customers. Any log-in process supporting this identification may also support the documentation of non-purchase decisions.

We recommend to impute missing data, as a complete data set leads to a better estimation of the underlying customer choices. When assuming that limited offer sets lead to non-purchase decisions and that a good imputation rule can yield approximate customer choices reasonably closely, imputation increases sample size compared to dropping censored observations. Therefore, imputing data enhances segmentation performance.

Here, we introduce four rules that impute missing data based on availability data. Additionally, we introduce two more sophisticated data imputation methods from the literature, k-nearest neighbours and multiple imputation with additive regression.

The four proposed imputation rules assume that products are ranked by an ordinal value, such as price or quality, so that product 1 has the lowest rank. The offer set at time t is characterized by the lowest-ranked product included,

$a_t \in S_t$, such that $x_t = a_t$ and all higher-ranked products are also included. When every product is offered, i.e., $a_t = 1$, no censoring results, as every customer can purchase a product from their consideration set.

When there is no natural order between products, applying the basic imputation rules presented in the following calls for creating an auxiliary ranking value from product attributes. When this is not feasible, other imputation methods are required. As examples for such more complex multiple imputation methods, we propose the *k*-nearest neighbours algorithm or multiple imputation methods.

The most basic rule, *none*, does not fill in missing observations. Instead, it introduces a new "no purchase" product, such that the dimension of matrix $\hat{P}$ increases by one. Now, each segment has a probability of non-purchase. Increasing the dimension of the matrix increases both computational effort and the number of potential segments.

The rule *lowest* replaces every no-purchase observation with a sale of the lowest-ranked product regardless whether this product was offered at the time of the observation. This rule may be applied when assuming that every customer would have at least bought the lowest product. Note that there is no missing data if all products were offered.

nextLower replaces missing data x_t^i with the next lower-ranked product compared to what the lowest-ranked product in the offer set, $a_t \in S_t$:

$$x_t^i = a_t - 1.$$

This assumes that customers were just barely not willing to purchase a product, e.g., given lacking willingness to pay.

The last two rules are based on the mean of the availability at time t and the lowest class. *meanFloor* defines

$$x_t^i = a_t - \left\lfloor \frac{a_t - 1}{2} \right\rfloor,$$

whereas *meanCeiling* defines

$$x_t^i = a_t - \left\lceil \frac{a_t - 1}{2} \right\rceil.$$

meanFloor and *meanCeiling* represent a compromize between *lowest* and *nextLower*. They only differ if the arithmetic mean $\frac{a_t - 1}{2}$ is not integer. Both assume that the customer would buy a product that is ranked roughly between the lowest-ranked product in the given offer set and the overall lowest-ranked product.

Finally, the rule *deleteObs* discards all observations which include missing data.

Table 2 exemplifies each rule. Here, the initial observation is $(7, -)$, where "$-$" indicates a non-purchase. For the first observed decision $t = 1$, $a_1 = 7$ was the lowest-ranked offered product, whilst at $t = 2$, it was $a_2 = 8$. As *none* does not account for the idea that customer choices depend on the offer set, it does not replace the observation. *lowest* replaces x_2^i with the lowest-ranked product 1, whilst *nextLower* fills the gap with $a_2 - 1 = 7$. *meanFloor* and *meanCeiling* replace it by $8 - \lfloor 3.5 \rfloor = 5$ and $8 - \lceil 3.5 \rceil = 4$, respectively. *deleteObs* discards the observation entirely and is therefore not represented in the table.

As an alternative to these simple rules, we also consider a *k-nearest neighbours* algorithm. *k*-nearest neighbours (*k-NN*) is a nonparametric approach to imputing missing data. It searches for the k nearest neighbouring data points according to a chosen metric. In our setting, we use the Euclidean metric and choose $k = \lfloor \sqrt{\bar{N}} \rfloor$, where $\bar{N}$ is the number of observations that are not non-purchases.

As another more sophisticated approach, we implement a *multiple imputation* (*MI*) method as described in Van Buuren (2012). Multiple imputation relies on three main steps: imputing missing data, analysing the completed data set, and aggregating results to form a final result. As the algorithm itself is relatively involved, we only describe it in outline and refer to Van Buuren (2012) for a comprehensive overview of multiple imputation.

The idea is to take several samples of the data set with missing values, predict those missing values, and finally combine the predictions for a final imputation. We implement the first step via bootstrapping the data set with missing values: For each variable with missing values, we create a sample with replacements from those observations where the variable is not missing. Then, we fit an additive model to the data set. This model is used to obtain predictions for the missing values in the original data set. We draw imputations from a multinomial distribution whose parameters are derived from the fitted models. These steps are done multiple times to account for uncertainty in the models. The final step combines the imputed values of the results into a final completed data set. To this end, we average over all results and take this mean, rounded to the next integer if necessary, as the imputed value.

Table 2 Example for missing data imputation

Observation		None	Lowest	NextLower	MeanFloor	MeanCeiling
$x_1^i = 7$	$a_1 = 7$	7	7	7	7	7
$x_2^i = -$	$a_2 = 8$	$-$	1	7	5	4

Results

As our primary practical contribution, we evaluate the performance of several variants of the proposed approach in practice-oriented settings. To evaluate segmentation accuracy, we simulate customer segments where customers choose between discrete products. For simulated data, the actual number of customer segments is known. This lets us assess the estimation accuracy of different model selection criteria and imputation methods by comparing the estimated rank to the actual number of segments.

We test different numbers of customer segments and different sample sizes. First, we analyse observations of discrete attributes and compare them to clustering indices. These results motivate a sensitivity analysis in section "Performance sensitivity to degrees of segment similarity", where we vary the overlap between two customers segments to analyse the performance sensitivity given structurally similar segments. Section "Performance on censored sales panel data" presents results from censored simulated sales panel data. Finally, we demonstrate the proposed approach's practical applicability on an empirical panel data set of airline bookings in section "Analysing an Empirical Data Sample of Airline Bookings".

If not stated otherwise, we report findings for 1000 stochastic samples of simulated data for an α-level of 0.05. Estimator performance is measured by the selection ratio of the correct number of segments, where we take $\hat{M}$ as the rank estimate of the corresponding model selection:

$$\text{Selection ratio} = \frac{\#\{\hat{M} = \text{actual number of segments}\}}{\#\{\text{stochastic samples}\}}$$

We processed all analyses in the statistics software R on a Quad-Core Intel Xeon CPU E3-1231 v3 3.40GHz with 8GB RAM and Windows 10. We employed the package *doParallel* by Revolution and Weston (2015) to use three out of four cores for computations. Depending on the number of observations per sample and dimension of the observed attributes, estimation over 1,000 stochastic samples took between 15 and 40 minutes. Full results with α-levels of $\{0.01, 0.05, 0.1\}$ are documented in "Supplementary results on estimation performance" section in Appendix.

Creating simulated observations

To evaluate the approach for observations of customer choice from ten discrete products D, we generate simulated observations as follows. First, we define M customer segments via their consideration set $D_i \subset D$. Consideration set $D_i \subset D$ indicates the subset of products segment i would be willing to buy. Next, we define the component weight vector

$\pi = (\pi_1, \ldots, \pi_M)$, which reflects the distribution of customers over segments. In addition, we set a total number N of individual customers to be observed.

The data creation process starts by drawing N realisations k of a random variable with distribution π, which represent the allocations of the current individual customer to a segment. Then, each observation is created by uniformly sampling two products with replacement from the consideration set D_k.

We define scenarios including three to six customer segments. Each segment's consideration set contains more than one product, i.e., $|D_i| > 1$. We identify products with the numbers 1 to 10. Segments are specified as follows:

$$D_1 = \{1, 2, 3, 4\}, \quad D_4 = \{5, 6, 7, 8\},$$
$$D_2 = \{3, 4, 5, 6\}, \quad D_5 = \{2, 3, 4, 5\},$$
$$D_3 = \{8, 9, 10\}, \quad D_6 = \{8, 9\}.$$

In addition, we generate a scenario with six segments, each buying a single product, i.e., $|D_i| = 1$. Each of the six segments considers exactly one of the six products on offer.

Example When creating observations for a market composed of equal parts of customers from segments D_1, D_2, and D_3, $\pi_1 = \pi_2 = \pi_3 = \frac{1}{3}$ ensures that each customer segment is equally likely to be selected. Allocation k is drawn from a uniform distribution and may turn out to be $k = 2$. Then, an observation from customer segment D_2 is created by uniformly drawing two samples from its consideration set $D_2 = \{3, 4, 5, 6\}$ with replacement. The result can be $(x_1, x_2) = (5, 3)$. This process is repeated until the desired number of observations has been created.

Estimation performance

For the first analysis, we consider three segments D_1, D_2, and D_3. We follow up by subsequently adding D_4, D_5, and D_6. We refer to the resulting demand scenarios by their highest index, D_3^* to D_6^*. Demand scenario D_6' includes six segments, where each consideration set includes exactly one product so that $D_i = \{i\}$. An overview of analysed scenarios is given in Table 3.

Table 3 Overview of scenarios with discrete attributes

Scenario	Customer segments
D_3^*	D_1, D_2, D_3
D_4^*	D_1, D_2, D_3, D_4
D_5^*	D_1, D_2, D_3, D_4, D_5
D_6^*	$D_1, D_2, D_3, D_4, D_5, D_6$
D_6'	$\{1\}, \{2\}, \{3\}, \{4\}, \{5\}, \{6\}$

The component weights are uniformly distributed across all segments, so that for M segments $\pi = (1/M, \ldots, 1/M)$. To assess estimation performance, we vary the sample size dependent on the number of segments. We report our findings for D_3^*, D_6^*, and D_6', as results for D_4^* and D_5^* are roughly similar. All further results can be found in "Supplementary results on estimation performance" section in Appendix.

Figure 1 shows the selection ratio of the correct number of segments for different sample sizes. Different line patterns mark the performance of the three model selection criteria, AIC, BIC, and SHT. The x-axis shows the sample size. Note that the gaps between data points are not equidistant for D_3^* and D_6^*. The y-axis displays the selection ratio. In the following, if a model selection criterion overestimates the number of segments, we speak of overfitting.

Figure 1a displays the performance of AIC, BIC, and SHT for D_3^*. Sample sizes include $\{100; 500; 1000; 5000; 10,000\}$. AIC performs comparatively well for small samples. However, it tends to overfit for large samples because it does not correct for the sample size. BIC outperforms SHT for all sample sizes up to 5000. Given 5000 or more observations, BIC and SHT both accurately estimate the number of segments.

Figure 1b displays the results for D_6^*. Sample sizes include $\{500; 1000; 5000; 10,000; 20,000; 50,000\}$. Again, AIC performs well for small samples but overfits for 5000 and more observations. BIC and SHT behave as observed for D_3^* in that both require larger sample sizes. BIC selects the correct number of segments in 95% of the cases for 20,000 observations. SHT needs more than 20,000 observations to reach a selection ratio higher than 90%.

In demand scenario D_6', the model selection criteria behave slightly differently. Sample sizes include $\{25; 50; 75; 100\}$. Across all methods, about 75 observations suffice to achieve a selection ratio of over 90%. Figure 1c shows that AIC performs best, as its tendency to overfit is beneficial in this case. Since the sample includes six products, the probability matrix $\hat{P}$ can at most have rank six. Hence, all estimates have an upper bound of six. Again, BIC outperforms SHT for small samples, but the gap closes at 75 observations.

To benchmark this approach, we evaluate four internal cluster validation measures with the common k-means clustering algorithm using the Euklidean metric: the Calinski-Harabasz index, the Silhouette index, the modified Hubert Γ statistic, and the S_{Dbw} validity index. When applied to the results of k-means given number of clusters k, each of these indices claims to point out the best number of clusters. In our terms, these correspond to the number of customer segments in the data.

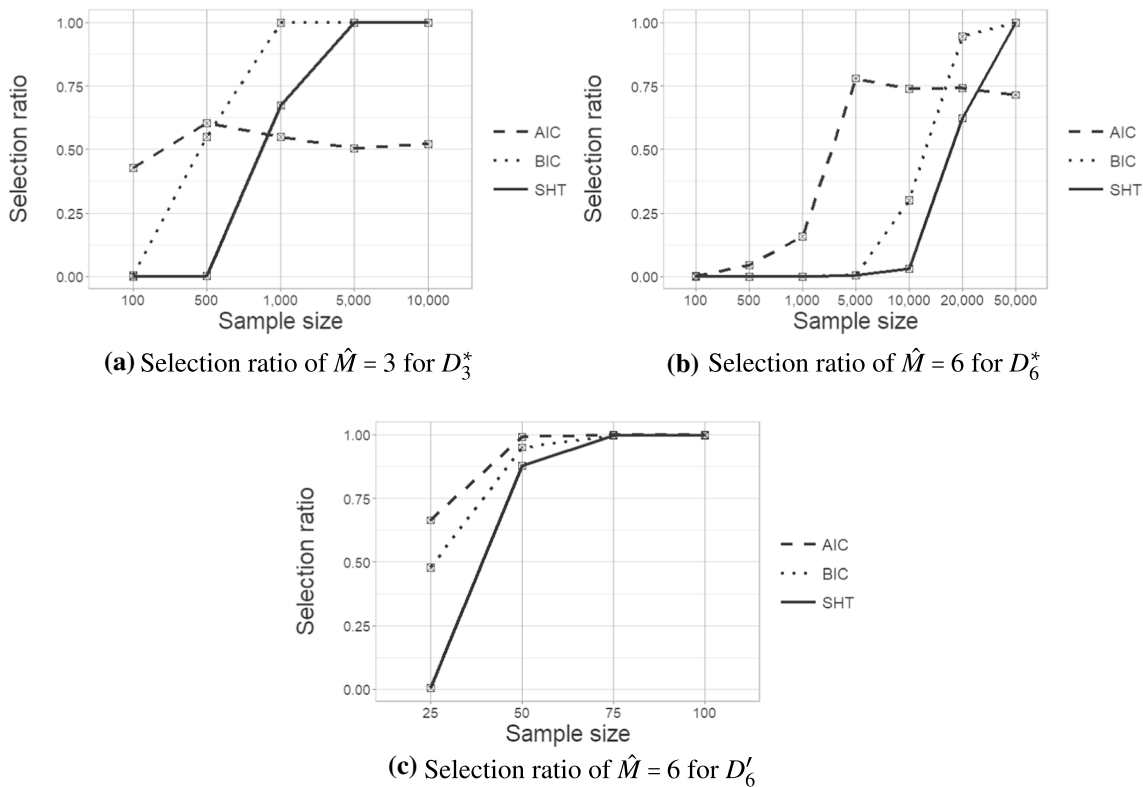The validation indices are based on two or three main properties of the data set: cluster compactness, cluster

(a) Selection ratio of $\hat{M} = 3$ for D_3^*

(b) Selection ratio of $\hat{M} = 6$ for D_6^*

(c) Selection ratio of $\hat{M} = 6$ for D_6'

Fig. 1 Discrete attributes—selection ratios for D_3^* (**a**), D_6^* (**b**), and D_6' (**c**)

separation, and cluster density. The *Calinski-Harabasz index* measures the ratio between separation and compactness with a normalization factor including the number of data points and the number of clusters (Caliński and Harabasz 1974). The *Silhouette index* is also based on compactness and separation, but averages these values over all single data points and the normalization is done via the maximum of compactness and separation (Rousseeuw 1987). The S_{Dbw} *index* evaluates clusters by compactness and density and uses the maximum of the density of clusters as normalization (Halkidi and Vazirgiannis 2001). The *modified Hubert* Γ *statistic* is a graphical method ('knee' or 'elbow criterion'), which calculates the distance between two data points and the corresponding cluster centres (Hubert and Arabie 1985).

Liu et al. (2010) discuss advantages and disadvantages of the individual indices. As the authors report, only S_{Dbw} performs well across all tested performance scenarios. Still, the remaining indices may perform well in suitable settings and are computationally faster than the S_{Dbw} index.

We limit the number of considered segments to the interval from two to eight. Applying these indices to the same data sets showed that only the *Calinski–Harabasz index* results in a non-zero selection ratio of the true underlying number of segments, as shown in Tables 4 and 5. Regardless of the number of underlying segments, the S_{Dbw} *index* chooses between six and eight segments with a high selection ratio of eight, the *Silhouette index* and the *modified Hubert* Γ *statistic* both always choose two segments.

Even though the *Calinski–Harabasz index* sometimes chooses the correct number of customer segments in scenario D_3^*, the selection ratio actually decreases with increasing sample size. This indicates that this index systematically underestimates the number of customer segments. With an increasing number of customer segments, the index performs even worse, differentiating only two segments. This is not surprising, as the customer segments partly overlap. This overlap lets the resulting observations form fewer distinct clusters than there are segments. k-means clustering cannot distinguish clusters that are not distinct. Since the cluster indices are based on the k-means clustering results, they also underestimate the number of segments. We conclude that traditional clustering with k-means does not yield good results in settings with overlapping customer segments. However, the proposed approach performs far better in these situations. In the following, we omit results from clustering indices when comparing selection ratios for different model selection criteria in Fig. 1.

These results motivate a sensitivity analysis to test how the proposed estimation algorithm performs when customer segments are more or less similar. We present results for such an analysis in the next section.

Performance sensitivity to degrees of segment similarity

We assess the performance sensitivity of the estimation approach when employing SHT to the similarity of customer segments by evaluating overlapping consideration sets. To that end, we simulate observations for two customer segments with varying degrees of overlap in their consideration sets. Again, we consider ten products $S = \{1, 2, 3, \ldots, 10\}$. In the most different case, for overlap 0, customer segment D_1 considers the first five products $D_1 = \{1, 2, 3, 4, 5\}$, whilst customer segment D_2 considers the last five products $D_2 = \{6, 7, 8, 9, 10\}$. We increase the overlap by alternately adding a product to each of the segments' consideration sets, i.e., overlap 1 adds product 6 to D_1, overlap 2 adds product 5 to D_2 and so on.

Figure 2 reports the selection ratios for $\hat{M} = 2$ and $\alpha = 0.05$. The x-axis shows the number of overlapping products; the y-axis shows the sample size. Darker shades

Table 4 Selection ratios for *Calinski–Harabasz index* for D_3^*

$\hat{M}$	Sample size				
	100	500	1000	5000	10,000
2	0.49	0.61	0.59	0.6	0.65
3	0.50	0.39	0.41	0.4	0.35
8	0.01	0	0	0	0

Table 5 Selection ratios for *Calinski–Harabasz index* for D_6^*

$\hat{M}$	Sample size			
	500	1000	5000	10,000
2	0.94	0.99	1	1
3	0.06	0.01	0	0
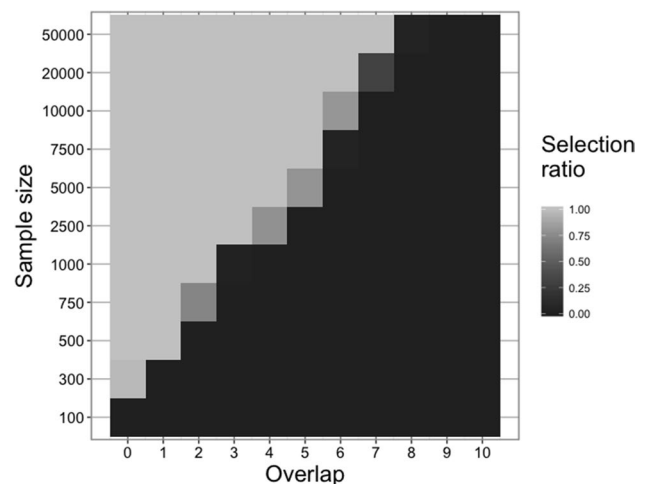
Fig. 2 Sensitivity analysis for two segments with increasingly overlapping consideration sets

indicate a low selection ratio, whereas lighter shades indicate a high selection ratio. Additional results can be found in "Discrete attributes" section in Appendix.

With increasing sample size, the estimation approach frequently differentiates the two segments, even when the overlap increases. Overlaps of 7 and more strongly increase the required sample size, as the probability matrix (3) becomes more and more homogeneous. Therefore, the estimation approach has a higher chance of estimating that there is a linear dependency within the matrix and thus it decreases the number of estimated segments to one. As expected, a full overlap of ten products renders D_1 and D_2 identical; for this case, the proposed approach should and does only identify a single customer segment. In "Continuous attributes" section in Appendix, we present a supplementary sensitivity analysis, focusing on the approach by Kasahara and Shimotsu (2014), which considers choices from alternatives defined by a common continuous product.

Performance on censored sales panel data

To assess estimation performance on censored sales panel data, we adapt the process of creating simulated observations. As before, we consider ten products and a uniform distribution of customers over M segments, i.e., $\pi = (1/M, \ldots, 1/M)$. Products are ranked on an ordinal scale, i.e., product 1 is the lowest-ranked and product 10 is the highest ranked product. The lowest-ranked product offered at time t, $a_t \in \{1, \ldots, 10\}$ defines the offer set, where all higher-ranked products are available. For each sales observation, we draw a lowest-ranked offered product from a uniform distribution across all products. Subsequently, observations are censored: If the observation includes only purchases of products ranked equal to or higher than that of the lowest-ranked offer, it is not modified. If the observation includes lower-ranked products than the lowest-ranked offer, the corresponding entry is deleted to create missing data, a.k.a. a non-purchase. This procedure censors approximately 71% of observations.

In the following, we report the results for censored versions of scenarios D_3^* and D_6^* (three and six customer segments). The rules *none*, *nextLower*, *meanFloor*, and *deleteObs* introduced in 3.4 and *k-NN*, *MI* are compared by their selection ratio given different sample sizes. The remaining results for are reported in "Supplementary results on performance on censored panel data" Appendix, as *lowest* yields results that are almost identical to the results of *none* and *meanCeiling* performs worse than *meanFloor*.

Figure 3 shows the selection ratio of the correct number of segments given different sample sizes. The selection ratios for BIC and SHT are shown in separate rows. We did not test AIC due to its overfitting behaviour in the previous results. The *x*-axis shows the estimated number of segments. The *y*-axis displays the selection ratio. The sample size is indicated by the shading of the bars.

Figure 3a shows the results for D_3^*. SHT and BIC perform comparably for up to 5000 samples. Whilst BIC accurately estimates the number of segments when using rules *none* and *lowest*, it underestimates for almost all other rules and sample sizes. Note that *none* essentially introduces a new "no purchase" product into the estimation procedure. Therefore, the resulting estimated numbers of segments are of questionable quality. We advise against not imputing censored data, but still report the corresponding findings by including the *none* rule.

Notably, *deleteObs* induces inconsistent results for BIC. BIC overfits the number of segments for more than 5000 observations. This can be explained by the process of creating simulated data. Since observations from customer segments that choose low ranked products are more prone to be censored, the estimation may recognize observations from these segments as belonging to additional segments. For example, when $x_1^i = 3$, the observation is censored by $a_1 > 3$ in 70% of all cases, whereas $x_1^i = 8$ is only censored by $a_1 \in \{9, 10\}$ in only 20% of all cases. Thus, segments that choose from low ranked products suffer more censoring, such that a linear dependence within the segment may disappear. The larger the sample, the clearer the segment's perceived differences become. Therefore, the number of segments is overestimated for large samples.

For SHT, the best selection ratios for large samples can be achieved by *meanFloor*. Analysing the simulated data showed that this approach reflects the choice of products in this data set, thus the rule performed best here. Whilst *deleteObs* performs worse in terms of selection ratio, it is still superior to the remaining rules with regard to large samples, especially given 71% censored observations. In this setting, censoring seems to be slightly detrimental to SHT with *none* and *lowest*, as the estimation approach tends to overfit for large samples. For smaller samples, rules *none* or *lowest* yield the best results.

The results show that *k-NN* is not suitable for the data set considered here, as it performs very poorly. Varying the weight to favour observations over availabilities does not significantly improve the performance of the *k*-nearest neighbours algorithm.

Both *k-NN* and *MI* overestimate the number of customer segments with increasing sample sizes. *k-NN* almost uniformly imputes observations over all products since it weights availability data, which is uniformly generated, equally to non-missing observations. This causes the algorithm to find those newly imputed customer segments and increase the estimated number of segments. It may be possible to find optimal weights for each demand setting but this would have to depend on the actual underlying segments. *MI* better represents customer segments, as it
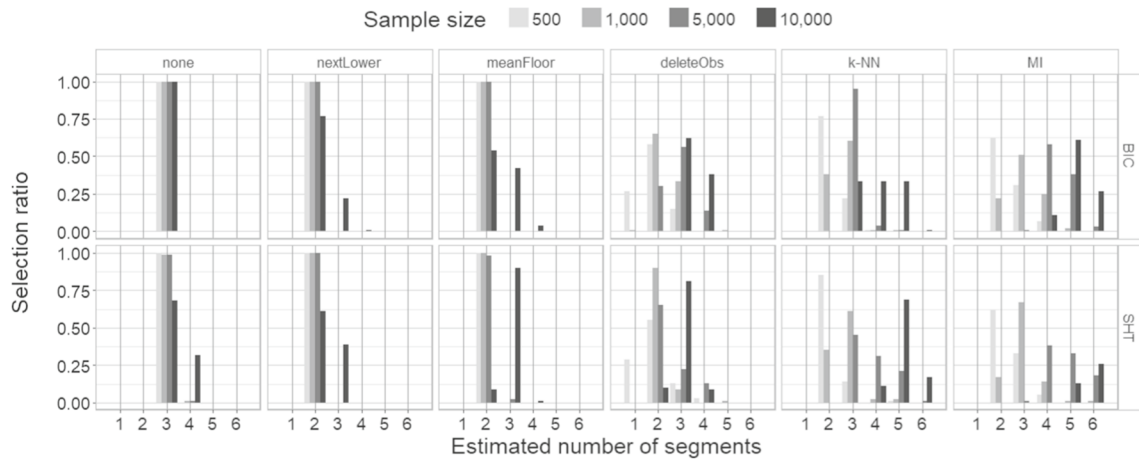
(a) Scenario D_3^*
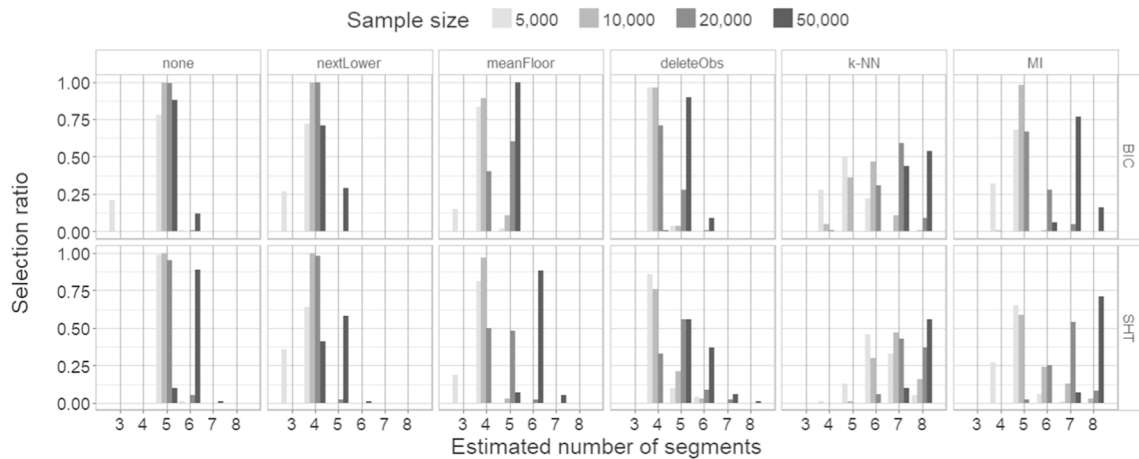
(b) Scenario D_6^*

Fig. 3 Selection ratios when imputing censored panel data

does not uniformly impute data. Still, since approximately 71% of all observations are censored, the bootstrapping procedure finds little variation in the data sets. Therefore, the predicted parameters are of questionable quality. This again causes flawed estimates.

Figure 3b depicts the selection ratios for D_6^*. In contrast to the results for uncensored observations, SHT outperforms BIC for all sample sizes and imputation rules. This is because BIC underestimates the number of segments, whereas SHT shows very promising results. rules *none*, *lowest* and *meanFloor* exhibit the best selection ratios with more than 80% for 50,000 observations. Again, *nextLower* causes underestimation, whilst *deleteObs* induces more varying results.

k-NN and *MI* produce similar results for D_6^* as for D_3^*. Both imputation methods do not correctly represent the underlying segments and overestimate of their number.

Additionally, we apply each imputation method 1000 times on different sample sizes to assess the computation

Table 6 Median of the computation time for data imputation methods in milliseconds

Method	$N = 500$	$N = 5,000$	$N = 20,000$	$N = 50,000$
nextLower	2.05	2.87	4.95	4.97
meanFloor	2.06	2.84	5.03	5.26
deleteObs	0.87	1.17	2.15	3.30
k-NN	7.94	92.62	1366.13	7559.17
MI	153.77	1450.63	14470.60	68175.76

time. We report the median in milliseconds in Table 6. Computation times increase slightly for the proposed rules, whilst increasing exponentially for *k-NN* and the multiple imputation method. Thus, we conclude that simple rules may be very beneficial to keep the computational effort low on larger data sets or practical settings with repeated forecasting or optimisation procedures.

Concluding, the estimation approach continues to perform well when sales data is censored. The sample size required to obtain good performance is similar to that for uncensored data. Surprisingly, in contrast to the results for uncensored data, BIC only rarely outperforms SHT. Generally, if customer behaviour can be replicated truthfully by imputation, both BIC and SHT are well-suited model selection criteria. Here, the rule *meanFloor* showed the best overall results. Given no assumptions about customer behaviour, *none* performed unexpectedly well.

Analysing an empirical data sample of airline bookings

To demonstrate practical applicability, we apply the estimation approach to an empirical sales panel data set of airline booking data. These data were collected via a frequent flier programme and was anonymized via replacing customer names by a unique identifier. The goal is to assess the number of segments within the domestic market based on the purchased booking classes per customer. In current practice, expert knowledge is used to estimate the number of customer segments in a market. The proposed approach increases reliance on empirical data, thereby contributing to circumvent subjective biases.

The empirical data set considered here contains a total of approximately 32,700 flight bookings. The sample includes the following attributes: customer identifier, date of purchase, date of departure, origin, destination, and booking class. It contains all bookings of domestic flights observed during one month in 2015. As opposed to groceries or drug stores, airline booking data is an example of scarce panel data.

As a preprocessing step, we remove all customers with less than two purchases. For odd numbers of purchases, one is randomly removed, for more than three purchases, we split the observed purchases into pairs. This creates a panel data set with 6733 individual observations of two purchase decisions. Even though the ratio of repeated purchases to total purchases is low, the proposed approach performs well.

The desired level of detail, at which the demand structure should be analysed, has to be considered when selecting the data set: The estimation approach strictly limits the analysis to the considered purchase decisions. For example, questions such as, "How many customer segments does the market of early morning (6–9 AM) flights between Frankfurt and
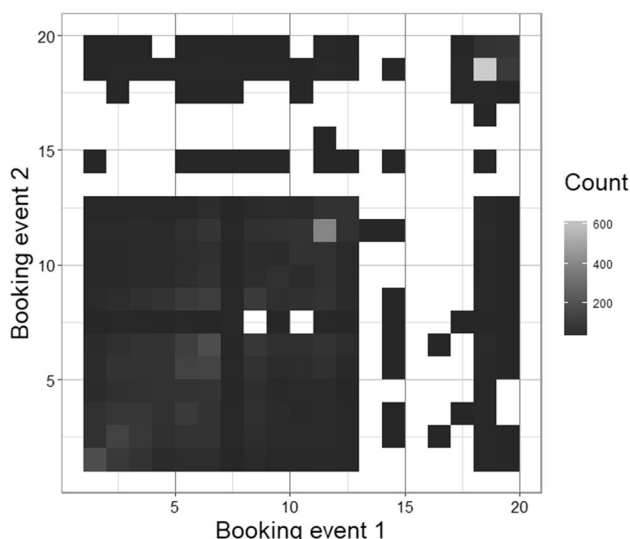

Fig. 4 Observation count of the empirical data sample

Munich on weekdays in summer contain?", can be answered by limiting the analysed data set accordingly, considering only ticket-purchases for flights that travelled from 6 to 9 AM between Frankfurt and Munich and during the summer.

In total, the data shows that customers booked one of 20 products are represented by booking classes. Thus, the upper bound for the number of segments is 20. Since empirical data is noisy, we expect more variance in estimation results than on simulated data.

Figure 4 visualizes the data set in terms of purchase events. The x-axis and y-axis show the booked class at time 1 and 2, respectively. Darker shades indicate fewer events, whereas lighter shades indicate more events. Deliberate blanks indicate there was no event at all. E.g., no customer was observed first booking product 12 and subsequently any product from 2 to 11 or 13 to 20. Supplementary results can be found "Supplementary results for the empirical data sample of airline bookings" section in Appendix.

Several factors influence the quality of empirical data. In the real world, customer choices are affected by many internal factors, such as willingness to pay and the reasons for travelling, and to external factors, such as product availability or competing offers. Therefore, the empirical data set is already censored by those factors. However, as data on such factors is not included in the data set, we do not know the actual degree of censoring. Additionally, different booking classes differ in price and product properties, such as being refundable, rebookable or offering a particular service level. Also, booking classes are not statically tied to a single price.

We do not know the actual number of customer segments that underlie the empirical data, precisely because it was not generated artificially. Therefore, analysing this data set can only demonstrate applicability in terms of computational

(a) Empirical data sample

(b) Artificially censored empirical data sample
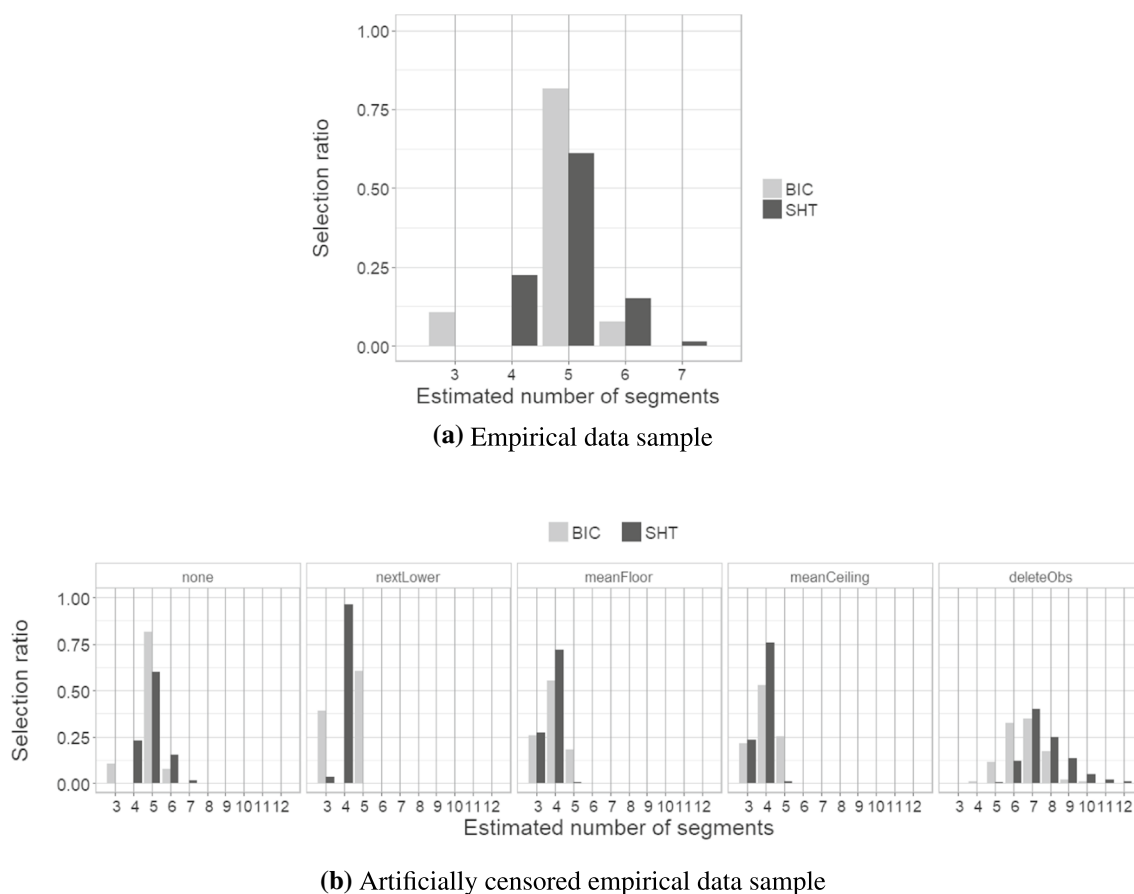
Fig. 5 Selection ratios for the empirical data sample without (**a**) and with (**b**) artificial censoring

effort and interpretable results. Since the empirical data does not contain any information about inherent censoring, such as, offer sets at the time of purchase, we do not know the extent of the censoring. Still, we expect most empirical observations to be already censored in some way. Nevertheless, we deliberately censor the data further to test the effect of imputation as described in section "Imputing censored sales data". Without information about offer sets, we would not be able to test the imputation rules. We apply the same censoring process as in section "Performance on censored sales panel data". The censored data set contains 4372 observations of which 1712 are uncensored by the drawn availabilities. Thus, we have a censoring rate of approximately 74%, which is comparable to the simulated data sets.

To create robust estimates, we used bootstrapping:[1] We resampled the data set 1000 times and report the selection ratios of BIC and SHT for α-level 0.05 in Fig. 5. Again, we did not test AIC due to its overfitting behaviour. Since the

previous results showed that the proposed method outperformed segmentation via clustering indices, and that simple rules outperform sophisticated imputation methods, we do not apply these alternatives to the empirical data set. The x-axes show the estimated number of segments. The y-axes display the selection ratio. Lighter grey bars indicate the performance of BIC, whilst darker grey bars show the performance of SHT.

Figure 5a depicts the selection ratios for 6,733 observations. BIC and SHT both exhibit the highest accurate selection ratio at five customer segments with 82% or 61%, respectively. The same effect can be seen in Fig. 5b with heuristic *none* and *lowest*. With *none*, BIC identifies five segments in 81%, whilst SHT does so in 60% of the cases. With *lowest* BIC identifies five segments in 84%, whilst SHT does so in 62% of the cases. The only difference in the highest selection ratio of BIC and SHT appears with *nextLower*. Here, SHT estimates four segments, whilst BIC estimates at five. For *meanFloor* and *meanCeiling*, both approaches estimate four customer segments. In contrast to the simulated data sets, the estimation performance using both heuristics is comparable. As before, *deleteObs* induces inconsistent results as its peaks are at a ratio less than 50%

---

[1] The concept of bootstraps was developed by Efron et al. (1979). It refers to increasing limited data sets by sampling new data sets from the original one. This is done via resampling with replacement.

and the number of estimated segments spans from 4 to 10 for BIC and from 6 to 12 for SHT.

Overall, the results exhibit more variance compared to the theoretical scenarios. This confirms our expectations for empirical data. Still, most approaches select four or five customer segments with high ratio, such that the estimation yields conclusive results for a given model selection criterion and heuristic.

Conclusion

This contribution proposed a nonparametric approach to estimate the number of segments from censored sales panel data of discrete products. We benchmarked three model selection criteria, sequential hypothesis testing (SHT), Akaike (AIC), and Baysian information criteria (BIC). In addition, we proposed and evaluated several approaches for imputing censored data.

The estimation approach employs a rank estimation of a probability matrix derived from sales panel data over two purchase decisions. The rank yields a lower bound for the number of customer segments present in a market, where segments were defined over their consideration sets of products.

An extensive benchmarking study based on simulated data sets helped to assess the performance of the approach in a controlled environment. We conclude that the algorithm outperforms estimating the number of segments by cluster quality indices calculated for the common k-means algorithm.

With regard to selecting the best model selection criterion, we found that BIC is best applied to discrete attributes, such as product purchases and booking classes, whilst SHT performs well for continuous attributes (see "Estimation performance for choices characterized by continuous attributes" section), such as prices. AIC overfitted the models for larger samples across data sets. We also analysed the sensitivity of estimation performance given different degrees of segment similarity.

As sales data can be expected to be censored, we presented four simple rules to impute missing observations. The numerical study highlights that these simple rules outperform more sophisticated imputation methods not only just in terms of computational effort, but also in terms of improving the segmentation algorithm's accuracy.

Finally, we applied the approach to an empirical airline booking data set and discussed its practical applicability.

Further research on this subject may consider a more sophisticated customer choice model to simulate the simulated data sets. Furthermore, the handling of availability data could be combined with more elaborate uncensoring methods to account for missing data points. An empirical data set with more consistent observations and non-artificial availability data could be employed to further assess the estimation performance in practical settings.

Appendix

Technical details

This appendix adds some technical results for the methodology presented in section "Methodology". Theorem 1 describes the asymptotic distribution of critical root statistic *CRT*.

Let $C = (c_1, \ldots, c_p)$ be a $p \times p$ matrix, where c_i is the eigenvector of matrix PP^T that belongs to the i-th largest eigenvalue λ_i and let $D = (d_1, \ldots, d_q)$ be the $q \times q$ matrix, where d_i is the eigenvector of matrix $P^T P$ that belongs to its i-th largest eigenvalue. Partition both C and D such that $C = (C_r, C_{p-r}) = (c_1, \ldots, c_r, c_{r+1}, \ldots, c_p)$ and $D = (D_r, D_{q-r}) = (d_1, \ldots, d_r, d_{r+1}, \ldots, d_q)$.

Theorem 1 (Robin and Smith (2000)) *Assume that*

(*i*) $\sqrt{n}\, vec(\hat{P} - P) \xrightarrow{d} \mathcal{N}(0, \Omega)$, where Ω is finite and of rank s, $0 < s \leq pq$,

(*ii*)   If $r_0 < q \leq p$, the matrix $(D_{q-r_0} \otimes C_{p-r_0})^T \Omega (D_{q-r_0} \otimes C_{p-r_0})$ is non-zero.

If $r_0 < q$, then statistic $CRT(r_0)$ (5) has an asymptotic distribution described by $\sum_{i=1}^{t} \gamma_i Z_i^2$, where $t \leq \min\{s, (p - r_0)(q - r_0)\}$, and $\gamma_1 \geq \ldots \geq \gamma_t$ are the non-zero ordered eigenvalues of the matrix $(D_{q-r_0} \otimes C_{p-r_0})^T \Omega (D_{q-r_0} \otimes C_{p-r_0})$, and $\{Z_i\}_{i=1}^{t}$ are independent standard normal random variables.

Here, $\xrightarrow{d}$ indicates the convergence in distribution and $A \otimes B$ denotes the outer or tensor product of matrix A with matrix B.

Proposition 1 establishes the convergence properties of the information criteria in section "Model selection".

Proposition 1 (Kasahara and Shimotsu (2014), *Proposition 6*) *Let*

$$\tilde{r} = \arg\,\min_{1 \leq r \leq q}\,(CRT(r) - f(N)g(r))$$

be the selected rank for an information criterion. If $f(N) \to \infty$, $f(N)/N \to 0$ *and* $P(g(r) - g(r_0) < 0) \to 1$ *for all* $r > r_0$ *as* $N \to \infty$, *then* $\tilde{r} \xrightarrow{p} r_0$.

Here, $\xrightarrow{p}$ indicates the convergence in probability.

Now, $g(r) = (p - r)(q - r)$ satisfies these conditions, if either $(d_r \otimes c_r)^T \Omega (d_r \otimes c_r) > 0$ for $1 \le r \le q$ or if there exists a pair (i, j) such that $(d_i \otimes c_j)^T \Omega (d_i \otimes c_j) > 0$ with $r + 1 \le i \le p$ and $r + 1 \le j \le q$ for any $1 \le r \le q$. Thus, for $g(r) = (p - r)(q - r)$ and $f(N) = 2$ or $f(N) = log(N)$, respectively, we have AIC or BIC, respectively. Note that the assumption $f(N) \to \infty$ is not satisfied by the AIC and therefore AIC does not estimate the right rank asymptotically.

Estimation from observations with continuous attributes

If we are interested in segmenting demand when observing continuous attributes, we may apply the original continuous version of Kasahara and Shimotsu (2014). To do so, we need to slightly change the way we count observations. In this setting, segments may be interpreted as groups of customers that purchase products in similar price categories or products with similar quality. Let x_t^i be observations with continuous attributes at time $t \in \{1, 2\}$. Since attributes are continuous, we cannot count the respective observations as individual elements, as in (4). To arrange observations into a matrix, we create partitions $I_1^1, I_2^1, \ldots, I_{n_1}^1$ and $I_1^2, I_2^2, \ldots, I_{n_2}^2$ with $n_1, n_2 \in \mathbb{N}^+$ such that

$$S_1 = \biguplus_{i=1}^{n_1} I_i^1 \text{ and } S_2 = \biguplus_{i=1}^{n_2} I_i^2, \text{ where } \biguplus$$

denotes the union of disjoint sets. Without loss of generality, choose $n_1 > n_2$. For example, when observing product prices, we may create non-overlapping price intervals, in which we count observations.

Similar to (4), we define the empirical probability as

$$\hat{P}(x_1 \in I_k^1, x_2 \in I_l^2) = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}_{(I_k^1, I_l^2)}(x_1^i, x_2^i),$$
$$\forall k \in \{1, \ldots, n_1\}, l \in \{1, \ldots, n_2\}. \quad (6)$$

Here, we use sample quantiles of observations to partition the continuous attribute space. The remaining estimation procedure is then performed identically to the discrete case.

Supplementary results on estimation performance

Here, we show a complete recollection of the results. All figures contain diagrams for α-levels 0.01, 0.05, and 0.1.

We report the full results for observations of discrete attributes complementing Fig. 1 in the main text. The diagrams show the performance of the three model selection approaches AIC (dashed line), BIC (dotted line), and SHT (solid line). The x-axes show the sample size. Note that sample sizes vary depending on the number of segments and the gaps between data points are not equidistant. The y-axis displays the selection ratio of the correct number of segments (Figs. 6, 7, 8).
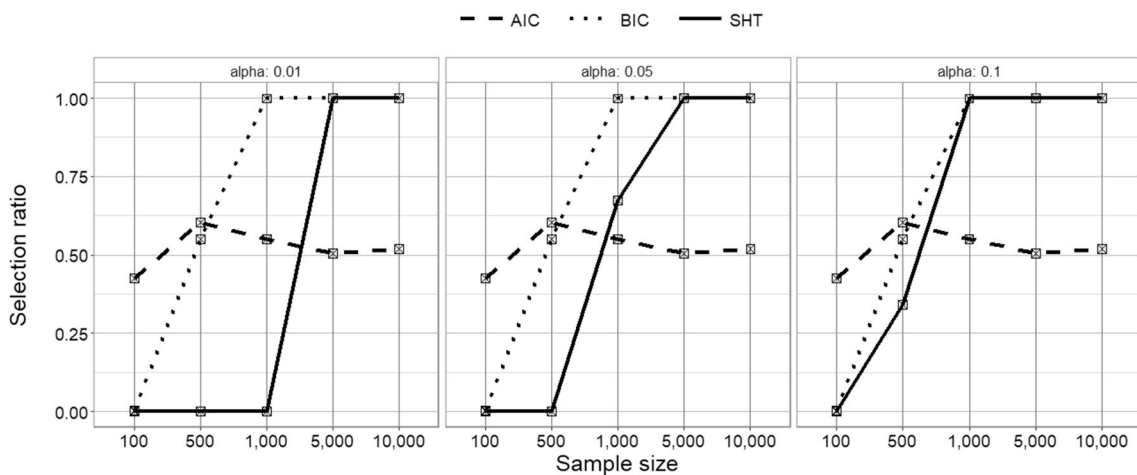

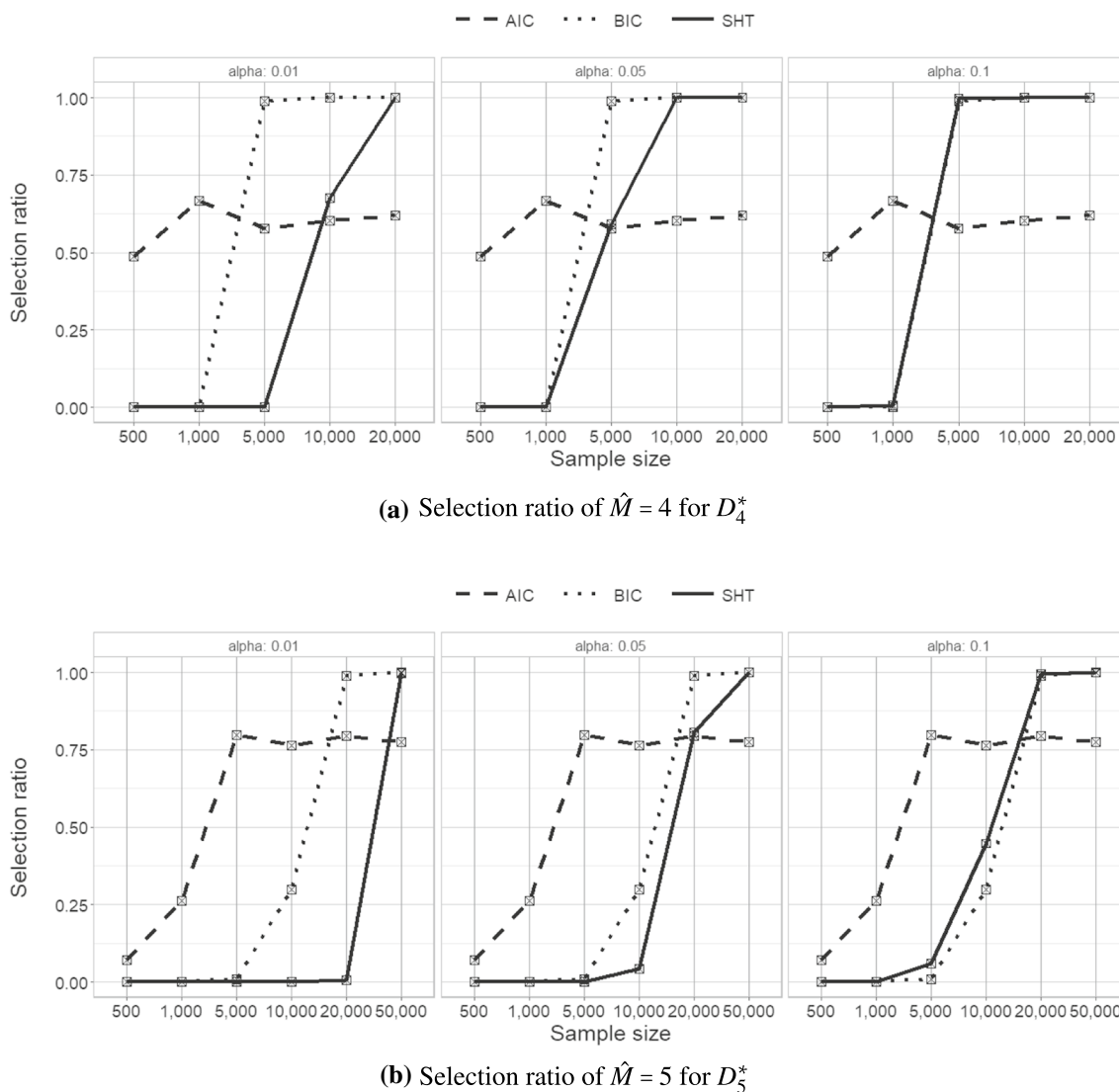**Fig. 6** Discrete attributes—selection ratio of $\hat{M} = 3$ for D_3^*

(a) Selection ratio of $\hat{M} = 4$ for D_4^*

(b) Selection ratio of $\hat{M} = 5$ for D_5^*

Fig. 7 Discrete attributes—selection ratios for D_4^* (**a**) and D_5^* (**b**)

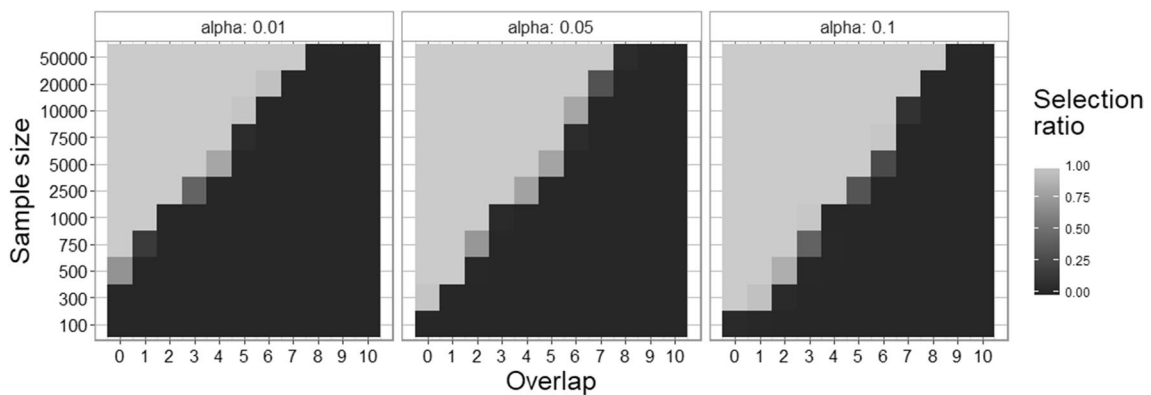## Supplementary results on sensitivity to segment similarity

Discrete attributes

This appendix reports results for different degrees of segment similarity complementing Fig. 2 in the main text. The results show selection ratios for $\hat{M} = 2$. The x-axis shows the overlap of the two customer segments, the y-axis shows the sample size. Darker shades indicate a low selection ratio, whereas lighter shades indicate a high selection ratio (Fig. 9).

Continuous attributes

This appendix reports results for different degrees of segment similarity complementing the discussion of continuous attributes "Estimation performance for choices characterized by continuous attributes" section in Appendix. The results show selection ratios for $\hat{M} = 2$. The x-axis shows the expected value μ_2^1, the y-axis shows the expected value μ_2^2. Values towards the lower left corner are closer to $\mu_1 = (0, 0)$. Darker shades indicate a low selection ratio, whereas lighter shades indicate a high selection ratio (Fig. 10).
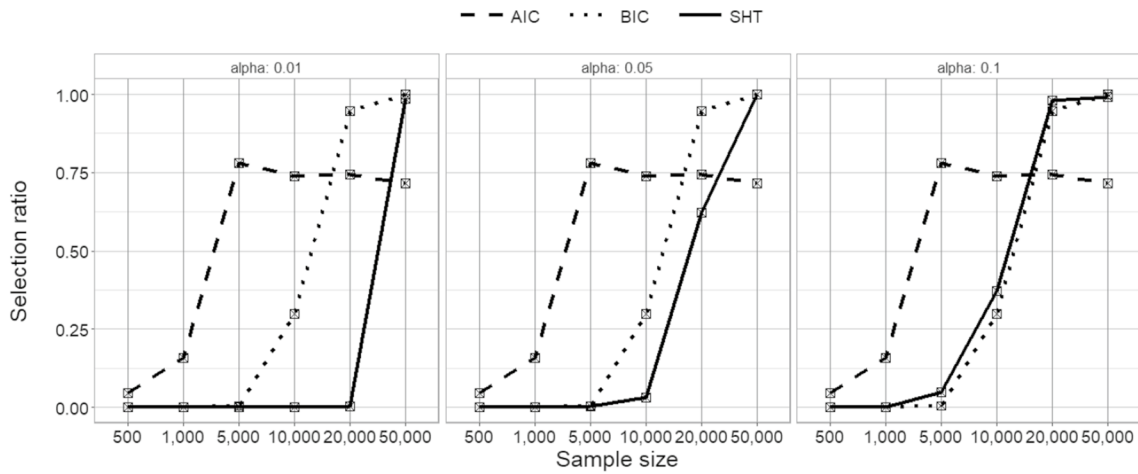
(a) Selection ratio of $\hat{M} = 5$ for D_5^*
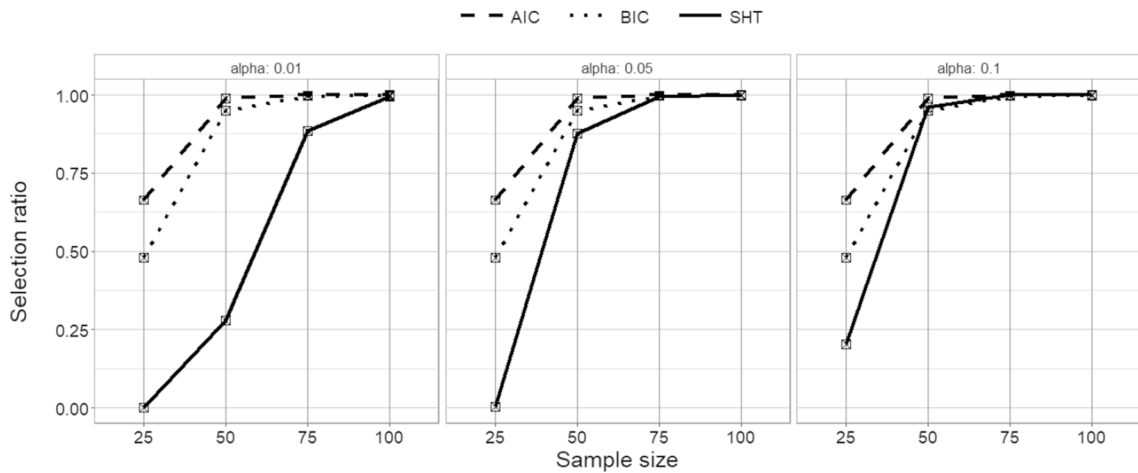
(b) Selection ratio of $\hat{M} = 6$ for D_6'

Fig. 8 Discrete attributes—selection ratios for D_6^* (**a**) and D_6' (**b**)

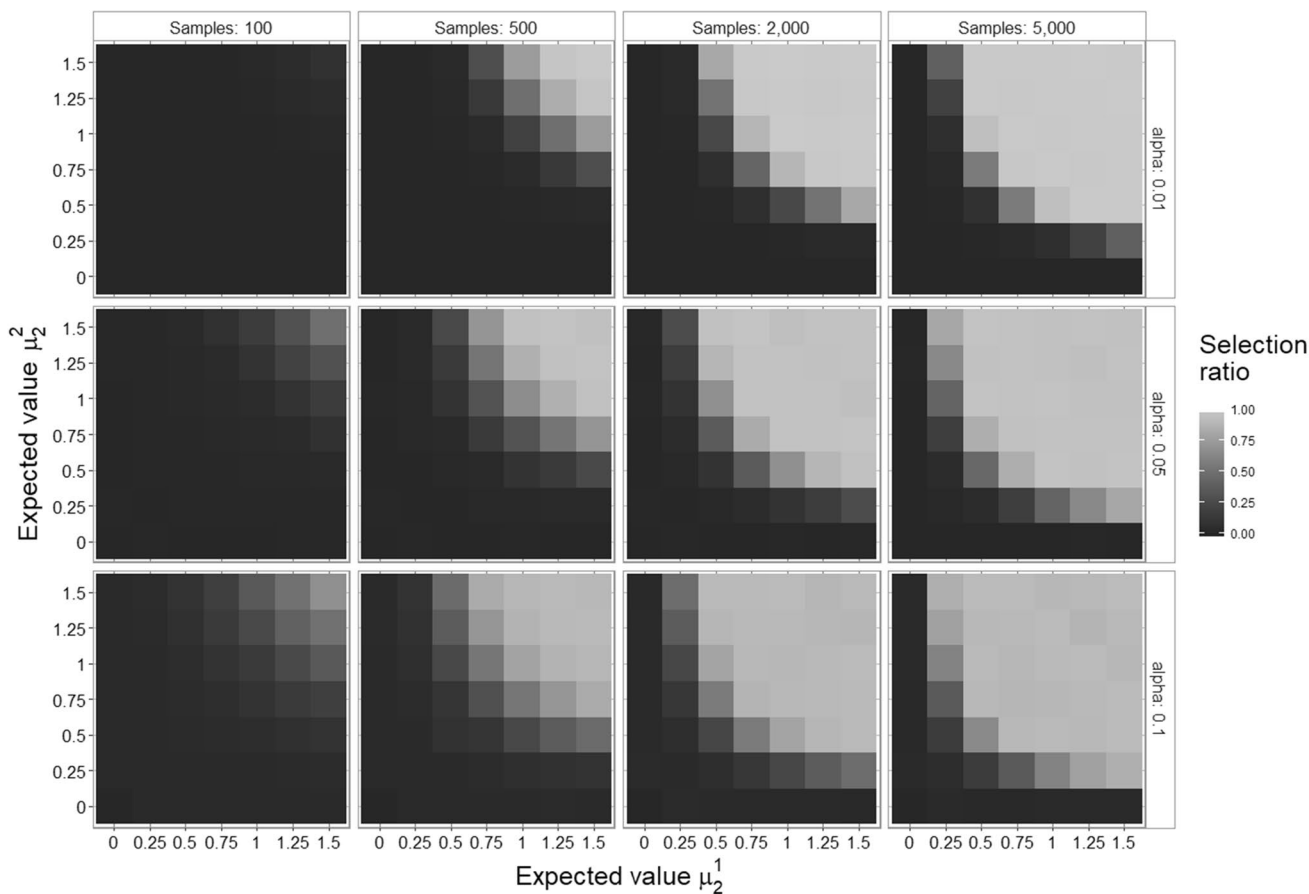**Fig. 9** Selection ratios for two segments with increasingly overlapping consideration sets

Fig. 10 Selection ratios for bivariate, normally distributed random variables

Supplementary results on performance on censored panel data

This appendix reports results for censored sales panel data of discrete attributes complementing Fig. 3 in the main text. The diagrams show the performance of the two model selection approaches BIC and SHT. The *x*-axes show the estimated number of segments. The *y*-axis displays the selection ratio. The sample size is indicated by the shading of the bars (Figs. 11, 12).

Estimation performance for choices characterized by continuous attributes

This appendix assesses estimation performance for observations differentiated by common continuous attributes *C*. In this case, we define customer segments through bivariate normal distributions, e.g., creating segments with a normally distributed willingness to pay. To validate the performance in other settings, we also analysed data sets with exponentially distributed random variables. On these,

the approach yields similar results to the results shown in this section.

Again, we consider scenarios with up to six segments, where the basic approach to generating simulated observations is similar to that for discrete attributes. We first draw the allocation to a segment and then a realisation of a bivariate, normally distributed random variable $X \sim \mathcal{N}_2(\mu, \Sigma)$, with segment-specific expected value μ and covariance matrix Σ. Here, we assume uncorrelated random variables with variance 1, i.e., $\Sigma = I_2$, where I_2 denotes the two-dimensional identity matrix. This variable defines the allocated segment. The segments are specified as follows:

$X_i \sim \mathcal{N}_2\big((\mu_i^1, \mu_i^2), \Sigma_i\big)$ with $\Sigma_i = I_2$ and

$\mu_1 = (0, 0),$	$\mu_4 = (5, 6),$
$\mu_2 = (2, 1),$	$\mu_5 = (-2, -3),$
$\mu_3 = (4, 3),$	$\mu_6 = (-1, 3).$

Note that $\mu_3^2 = \mu_6^2$, such that these two segments generate similar data for the second purchase decision. This is deliberately chosen to assess the behaviour of the estimation
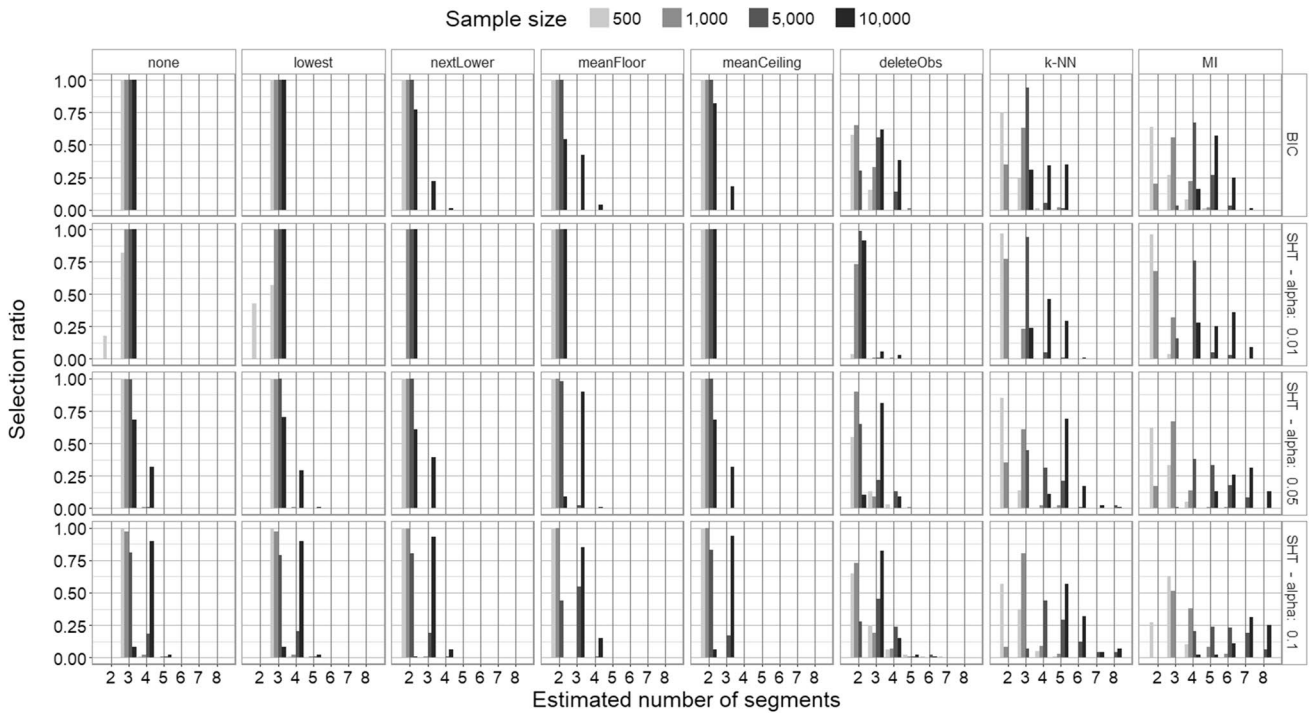
Fig. 11 Selection ratios for heuristics with censored panel data and three customer segments
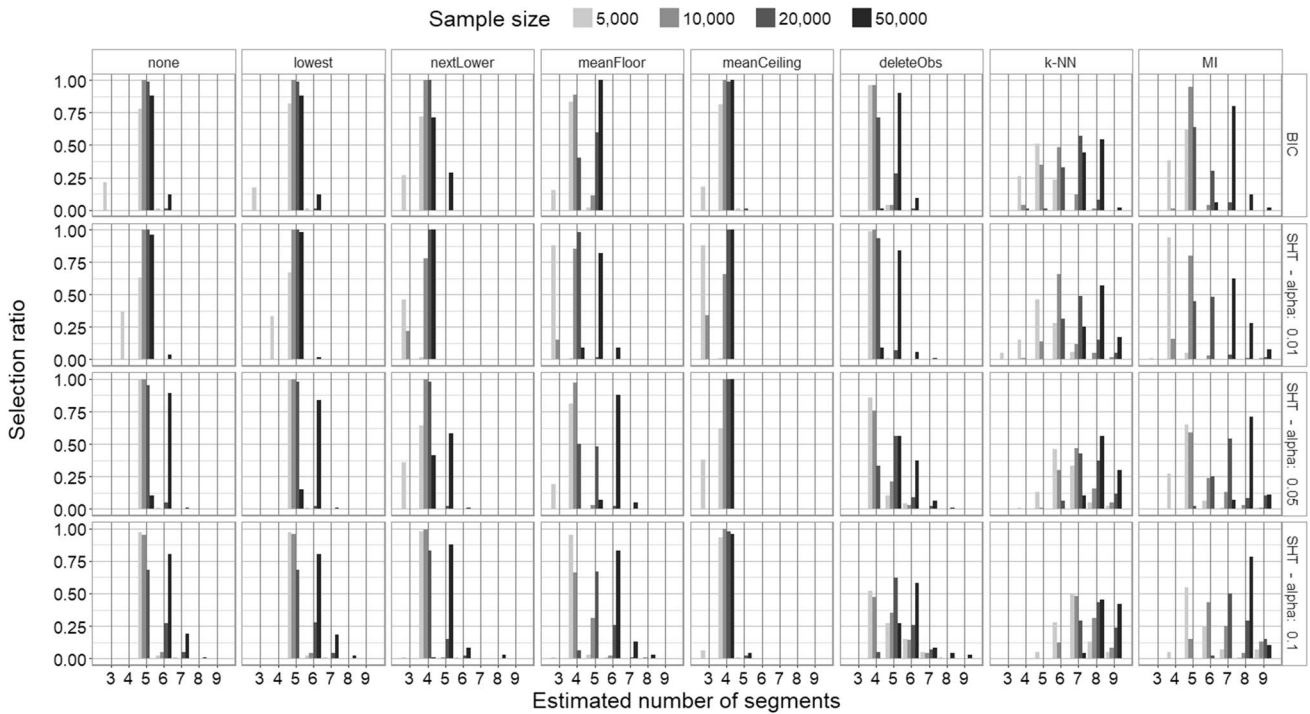

Fig. 12 Selection ratios for heuristics with censored panel data and six customer segments

Table 7 Overview of scenarios with continuous attributes

Scenario	Customer segments
C_3	μ_1, μ_2, μ_3
C_4	$\mu_1, \mu_2, \mu_3, \mu_4$
C_5	$\mu_1, \mu_2, \mu_3, \mu_4, \mu_5$
C_6	$\mu_1, \mu_2, \mu_3, \mu_4, \mu_5, \mu_6$

approach in such a case. We create data sets from scenarios including three to six customer segments and refer to these scenarios by C_3 to C_6. Table 7 lists all analysed scenarios. Again, the number of observations varies per data set, depending on the scenario. As previously, we report our findings only for C_3 and C_6. We omit any further discussion of the results from the benchmarked cluster indices as given in section "Estimation performance", limiting ourselves to the conclusion that the findings from applying these indices to continuous data are similar to those from applying them to discrete data.

Figure 13 shows the selection ratio of the correct number of segments for different sample sizes. Different line patterns mark the performance of the model selection criteria AIC, BIC, and SHT. The x-axis shows the sample size. Note that the gaps between data points are not equidistant. The y-axis displays the selection ratio.

Figure 13a shows the selection ratio of $\hat{M} = 3$ for C_3. Sample sizes include $\{100; 500; 1000; 5000; 10,000\}$. AIC

performs relatively well for small samples, but exhibits its known overfitting for larger samples. SHT outperforms BIC for smaller samples and is slightly less accurate for larger samples. Overall, we can deduce that for three customer segments, 1000 individual observations suffice to accurately estimate the number of segments with both BIC and SHT. At this point, the selection ratio for SHT is 91%, whilst BIC selects three segments in 86% of cases.

Because μ_3^2 and μ_6^2 are identical in C_6, the empirical probability matrix $\hat{P}$ (6) counts these observations in the same row or column. This results in a linear dependence within the matrix and a decreasing rank. Thus, segments 3 and 6 are expressed as a single segment in the finite mixture model. Therefore, Fig. 13b shows the selection ratio of $\hat{M} = 5$ for C_6. Sample sizes include $\{100; 500; 1000; 5000; 10,000; 20,000\}$.

Again, AIC performs comparatively well for small samples, but overfits the number of segments for more than 5000 observations. Similar to C_3, SHT slightly outperforms BIC for smaller samples and is slightly less accurate for larger samples. Overall, their performance is comparable for these settings. The analysis suggests that for five segments, 5000 observations suffice to obtain reasonable estimates. For that sample size, SHT selects five segments in 98% of the samples, whereas BIC exhibits a selection ratio of 95%.
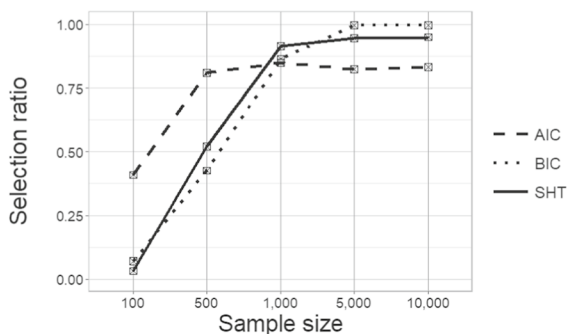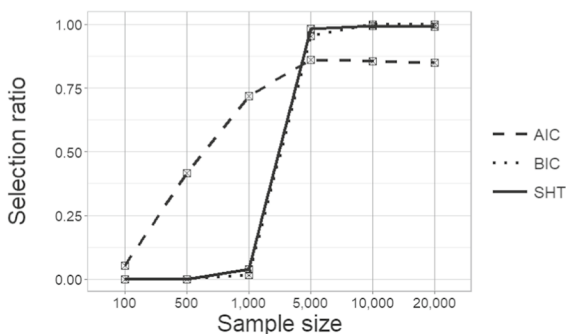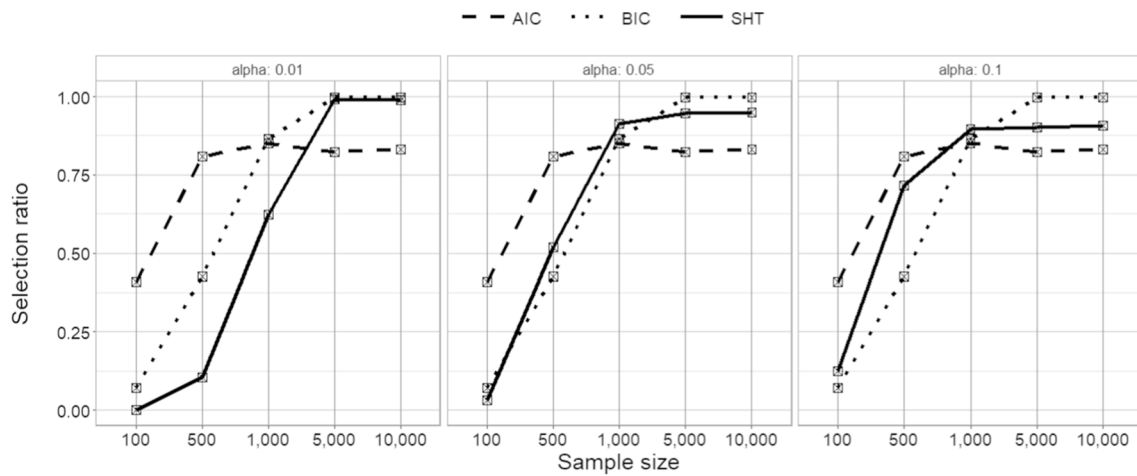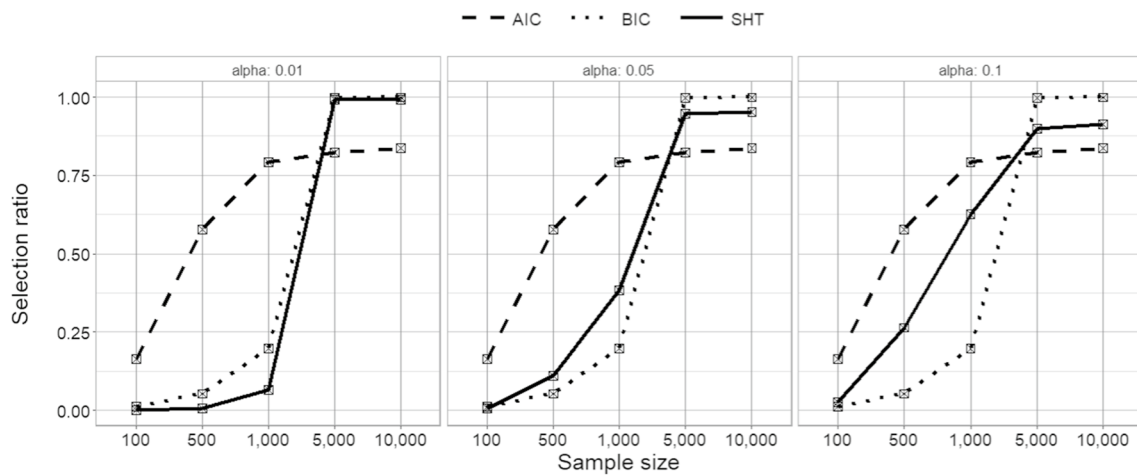

(a) Selection ratio of $\hat{M} = 3$ for C_3 (b) Selection ratio of $\hat{M} = 5$ for C_6

Fig. 13 Continuous attributes—selection ratios for C_3 (**a**) and C_6 (**b**)
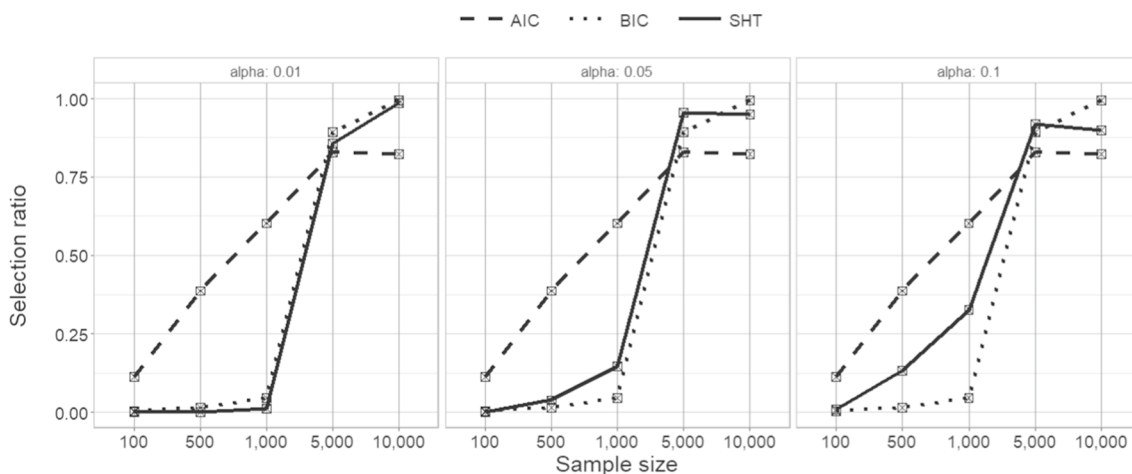
(a) Selection ratio of $\hat{M} = 3$ for C_3^*

(b) Selection ratio of $\hat{M} = 4$ for C_4^*

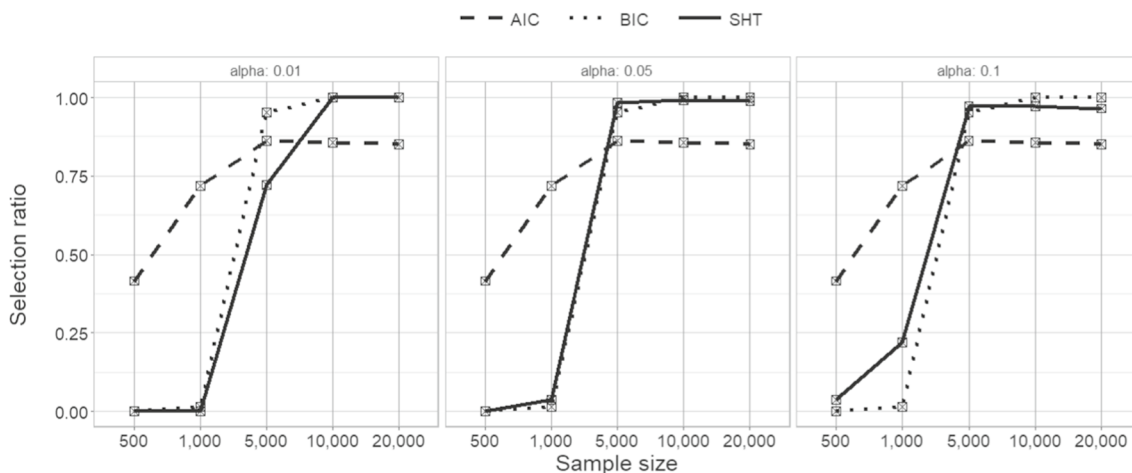**Fig. 14** Continuous attributes—selection ratios for C_3^* (**a**) and C_4^* (**b**)

In the following, we also report the results for the remaining α-values. Figures 14 and 15 show the performance of the three model selection approaches AIC (dashed line), BIC (dotted line), and SHT (solid line). The x-axes show the sample size. Note that sample sizes vary depending on the number of segments and the gaps between data points are not equidistant. The y-axis displays the selection ratio of the correct number of segments.

(a) Selection ratio of $\hat{M} = 5$ for C_5^*

(b) Selection ratio of $\hat{M} = 5$ for C_6^*

Fig. 15 Continuous attributes—selection ratios for C_5^* (**a**) and C_6^* (**b**)

Supplementary results for the empirical data sample of airline bookings

This appendix reports results for the empirical data set complementing Fig. 5 in the main text. We resampled the data set 1000 times and report the selection ratios of BIC and SHT. The x-axes show the estimated number of segments. The y-axes display the selection ratio. The lighter grey bars indicate the performance of BIC, whilst the darker grey bars show the performance of SHT. AIC is not represented due to its overfitting behaviour (Figs. 16, 17).
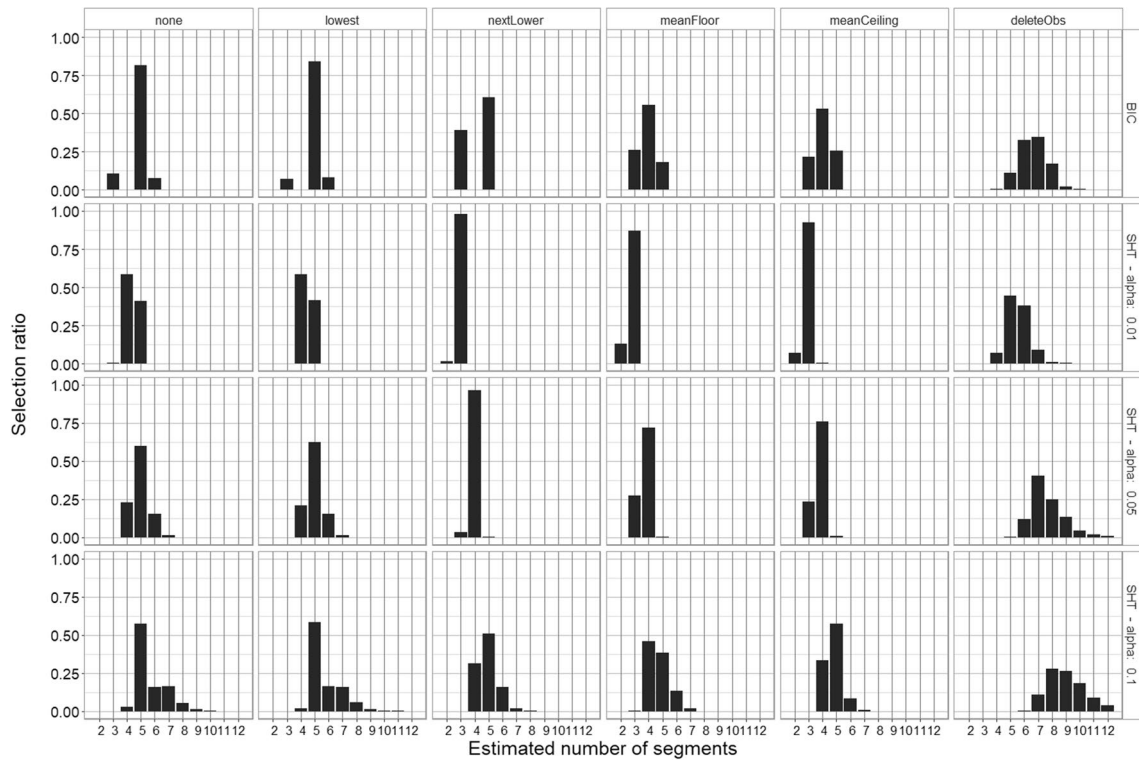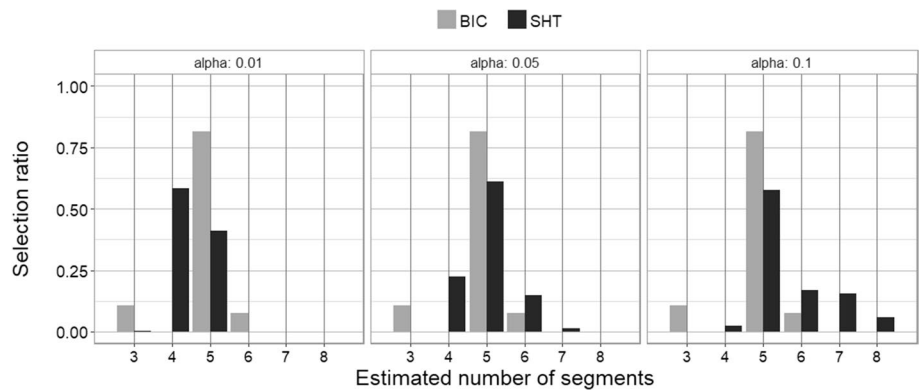
Fig. 16 Artificially censored empirical data sample

Fig. 17 Empirical data sample

References

Agresti, Alan, Brian Caffo, and Pamela Ohman-Strickland. 2004. Examples in which misspecification of a random effects distribution reduces efficiency, and possible remedies. *Computational Statistics & Data Analysis* 47 (3): 639–653.

Azadeh, Shadi Sharif. 2013. *Demand Forecasting in Revenue Management Systems*. École Polytechnique de Montréal: PhD diss.

Azadeh, Shadi Sharif, M. Hosseinalifam, and G. Savard. 2015. The impact of customer behavior models on revenue management systems. *Computational Management Science* 12 (1): 99–109.

Azadeh, Shadi Sharif, P. Marcotte, and G. Savard. 2015. A non-parametric approach to demand forecasting in revenue management. *Computers & Operations Research* 63: 23–31.

Bronnenberg, Bart J., Michael W. Kruger, and Carl F. Mela. 2008. Database paper-The IRI marketing data set. *Marketing science* 27 (4): 745–748.

Caliński, Tadeusz, and Jerzy Harabasz. 1974. A dendrite method for cluster analysis. *Communications in Statistics-theory and Methods* 3 (1): 1–27.

Cohen, Joel E., and Uriel G. Rothblum. 1993. Nonnegative ranks, decompositions, and factorizations of nonnegative matrices. *Linear Algebra and its Applications* 190: 149–168.

Efron, B., et al. 1979. Bootstrap Methods: Another Look at the Jackknife. *The Annals of Statistics* 7 (1): 1–26.

Farias, Vivek F., Srikanth Jagabathula, and Devavrat Shah. 2013. A nonparametric approach to modeling choice with limited data. *Management Science* 59 (2): 305–322.

Haensel, Alwin, and Ger Koole. 2011. Estimating unconstrained demand rate functions using customer choice sets. *Journal of Revenue & Pricing Management* 10 (5): 438–454.

Halkidi, Maria, and Michalis Vazirgiannis. 2001. "Clustering validity assessment: Finding the optimal partitioning of a data set." In Data Mining, 2001. ICDM 2001, Proceedings IEEE International Conference on, 187–194. IEEE.

Härdle, Wolfgang, and Enno Mammen. 1993. Comparing nonparametric versus parametric regression fits. *The Annals of Statistics*, 1926–1947.

Hubert, Lawrence, and Phipps Arabie. 1985. Comparing partitions. *Journal of Classification* 2 (1): 193–218.

Jagabathula, Srikanth, and Gustavo Vulcano. 2015. "A Model to Estimate Individual Preferences Using Panel Data." Available at SSRN 2560994.

Kasahara, Hiroyuki, and Katsumi Shimotsu. 2014. Non-parametric identification and estimation of the number of components in multivariate mixtures. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 76 (1): 97–111.

Kunnumkal, Sumit. 2014. Randomization approaches for network revenue management with customer choice behavior. *Production and Operations Management* 23 (9): 1617–1633.

Linoff, Gordon S, and Michael JA Berry. 2011. Data mining techniques: for marketing, sales, and customer relationship management. John Wiley & Sons.

Liti'ere, Saskia, Ariel Alonso, and Geert Molenberghs. . 2008. The impact of a misspecified random-effects distribution on the estimation and the performance of inferential procedures in generalized linear mixed models. *Statistics in medicine* 27 (16): 3125–3144.

Little, Roderick JA., and Donald B. Rubin. 2014. *Statistical analysis with missing data*. Hoboken: Wiley.

Liu, Yanchi, Zhongmou Li, Hui Xiong, Xuedong Gao, and Junjie Wu. 2010. "Understanding of internal clustering validation measures." In Data Mining (ICDM), 2010 IEEE 10th International Conference on, 911–916. IEEE.

McLachlan, Geoffrey, and David Peel. 2000. *Finite mixture models*. Hoboken: Wiley.

Meissner, Joern, Arne Strauss, and Kalyan Talluri. 2013. An enhanced concave program relaxation for choice network revenue management. *Production and Operations Management* 22 (1): 71–87.

Moe, Wendy W., and Peter S. Fader. 2004. Dynamic conversion behavior at e-commerce sites. *Management Science* 50 (3): 326–335.

Müller, Sven, and Knut Haase. 2014. Customer segmentation in retail facility location planning. *Business Research* 7 (2): 235–261.

Queenan, Carrie Crystal, Mark Ferguson, Jon Higbie, and Rohit Kapoor. 2007. A comparison of unconstraining methods to improve revenue management systems. *Production and Operations Management* 16 (6): 729–746.

Ratliff, Richard M., B. Venkateshwara Rao, Chittur P. Narayan, and Kartik Yellepeddi. 2008. A multi-flight recapture heuristic for estimating unconstrained demand from airline bookings. *Journal of Revenue and Pricing Management* 7 (2): 153–171.

Revolution Analytics and Steve Weston. 2015. doParallel: Foreach Parallel Adaptor for the 'parallel' Package. R package version 1.0.10. https://CRAN.R-project.org/packag e=doParallel.

Robin, Jean-Marc., and Richard J. Smith. 2000. Tests of rank. *Econometric Theory* 16 (02): 151–175.

Rousseeuw, Peter J. 1987. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics* 20: 53–65.

Rusmevichientong, Paat, David Shmoys, Chaoxu Tong, and Huseyin Topaloglu. 2014. Assortment optimization under the multinomial logit model with random choice parameters. *Production and Operations Management* 23 (11): 2023–2039.

Salch, J. 1997. Unconstraining passenger demand using the EM algorithm. In Proceedings of the INFORMS Conference.

Saleh, R. 1997. Estimating lost demand with imperfect availability indicators. In AGIFORSReservations and Yield Management Study Group Meeting Proceedings. Montreal, Canada.

Skwarek, Daniel Kew. 1996. Competitive impacts of yield management system components: forecasting and sell-up models. Master's thesis. Cambridge, Mass.: Massachusetts Institute of Technology, Flight Transportation Laboratory.

Talluri, Kalyan. 2009. A finite-population revenue management model and a risk-ratio procedure for the joint estimation of population size and parameters. Available at SSRN 1374853.

Talluri, Kalyan T., and Garrett J. Van Ryzin. 2004. *The theory and practice of revenue management*, vol. 68. New York: Springer.

Van Buuren, Stef. 2012. *Flexible imputation of missing data*. Boca Raton: CRC Press.

Van Ryzin, Garrett, and Gustavo Vulcano. 2015. A market discovery algorithm to estimate a general class of nonparametric choice models. *Management Science* 61 (2): 281–300.

Weatherford, Larry R., and Stefan Pölt. 2002. Better unconstraining of airline demand data in revenue management systems for improved forecast accuracy and greater revenues. *Journal of Revenue and Pricing Management* 1 (3): 234–254.

Johannes F. Jörg started his doctoral research at the Advanced Analytics Department of RWTH Aachen University in cooperation with Deutsche Lufthansa AG. He is mainly interested in nonparametric statistics, demand estimation and simulations in revenue management. He uses demand estimation procedures to divulge demand segments from observed data to gain insights into market structures. In simulations, he uses these methods to facilitate the calibration of demand and thus increasing their efficiency.

Catherine Cleophas From 2013 to 2017, Professor Catherine Cleophas headed the Research Group Advanced Analytics in the research area Operations Research and Management at RWTH Aachen University. From 2017, she held a position as Senior Lecturer Operations Research at Lancaster University Management School. Since October 2018, she chairs the research group Service Analytics at Christian-Albrechts-Universität zu Kiel. Her research interests include demand analytics and management, system design to support and complement human decision-making and agent-based simulations for business planning.