Heger, Julia; Klein, Robert

Article — Published Version

Assortment optimization: a systematic literature review

OR Spectrum

Suggested Citation: Heger, Julia; Klein, Robert (2024) : Assortment optimization: a systematic literature review, OR Spectrum, ISSN 1436-6304, Springer Berlin Heidelberg, Berlin/Heidelberg, Vol. 46, Iss. 4, pp. 1099-1161, https://doi.org/10.1007/s00291-024-00752-4

This Version is available at:
https://hdl.handle.net/10419/313826

OR Spectrum (2024) 46:1099–1161
https://doi.org/10.1007/s00291-024-00752-4

ORIGINAL ARTICLE

Assortment optimization: a systematic literature review

Julia Heger[1] · **Robert Klein**[1]

Received: 16 June 2023 / Accepted: 29 January 2024 / Published online: 16 April 2024

Abstract
Assortment optimization is a core topic of demand management that finds application in a broad set of different areas including retail, airline, hotel, and transportation industries as well as in the healthcare sector. Hence, the interest in research on assortment optimization has grown rapidly in recent years. However, the sheer number of publications on the topic of assortment optimization makes it difficult to keep track of all available approaches proposed in the literature. In this paper, we systematically review state-of-the-art studies on assortment optimization. We assemble an extensive literature overview by strategically searching for pre-defined keywords within leading scientific databases. The resulting literature is grouped by a proposed taxonomy that captures properties related to the optimization problem itself, the modelled customer behaviour, and the solution concept applied for solving the problem at hand. For each group, we provide an overview of the corresponding literature and analyse it based on a proposed selection of key factors.

Keywords Choice modelling · Assortment optimization · Parametric · Nonparametric · Review

1 Introduction

Assortment optimization is a core problem that arises in disciplines such as retail operations or revenue management (Qi et al. 2020) and finds application in a broad set of different areas including retail, airline, hotel, and transportation industries as well as in the healthcare sector. The assortment problem involves a seller choosing an appropriate subset of items from the available universe to be offered to a group of customers to maximize an objective, e.g. the expected revenue, while accounting for

✉ Robert Klein
robert.klein@uni-a.de

Julia Heger
julia.heger@uni-a.de

[1] Chair of Analytics & Optimization, University of Augsburg, Universitätsstr. 16, 86159 Augsburg, Germany

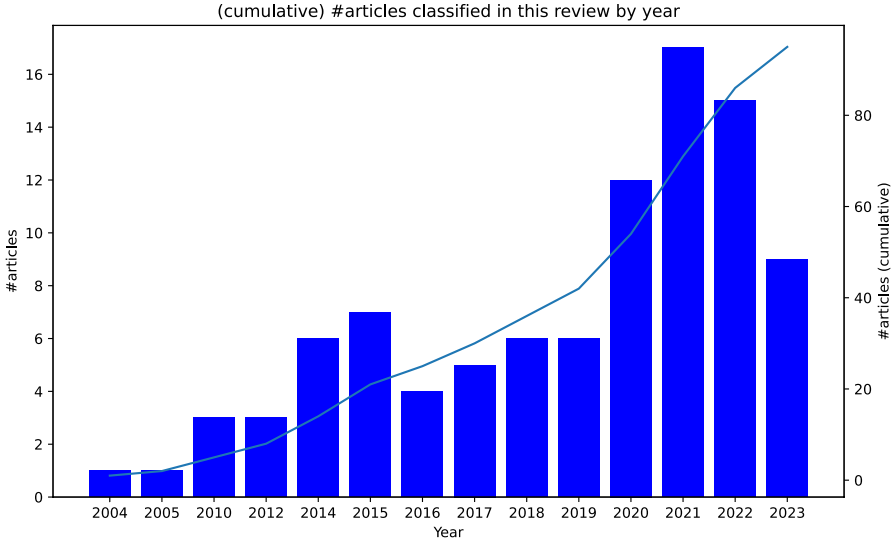

Fig. 1 Timely development of the (cumulative) number of journal articles related to assortment optimization that are classified in our literature review

the customers' choice behaviour. On the other hand, customers decide whether and which of the offered item(s) to purchase based on their preferences (Qi et al. 2020; Mišić and Perakis 2019). This results in a combinatorial problem that is extensively studied in the literature (Qi et al. 2020). Particularly over the last decades, assortment optimization received a considerable boost in attention both from practitioners and academics alike and became a highly active research area. This is also reflected by the development of the number of published journal articles related to this research direction. As can be seen in Fig. 1, the amount of articles related to assortment optimization published per year that are classified in our review increased strongly in recent years.[1]

Due to the sheer amount of literature on assortment optimization, it might be difficult to keep track of all available approaches. Surveys serve the purpose of providing a comprehensive overview of the most important publications, their approaches, and their findings.

There exist a variety of surveys covering the topic of assortment optimization. However, these surveys are often dedicated to a broader research area such as revenue management or retail operations and thus only briefly consider assortment optimization as an individual topic.

For example, the survey of Mišić and Perakis (2019) reviews applications of data analytics in operations management for three main areas, namely supply chain management, revenue management, and healthcare operations. As part of

[1] Note that the strong dip in 2023 results from the fact that this literature review only contains articles that are published until including April 2023.

the survey on revenue management, the authors provide a brief review of choice modelling and assortment optimization by presenting the findings of selected studies published in this area.

Strauss et al. (2018) conduct a survey on choice-based revenue management. They focus on the design and estimation of discrete choice models for revenue management and on the dynamic availability control problem under customer choice behaviour, which contains an assortment optimization problem. Hence, the authors provide a brief overview of related literature on assortment optimization, which is structured according to the underlying choice models.

In contrast, Qi et al. (2020) review literature on data-driven research in retail operations with a particular focus on studies in three core aspects of retail operations, namely assortment optimization, order fulfillment, and inventory management. The section on assortment optimization starts with an introduction to parametric choice models and their estimation, followed by the description of selected literature on assortment optimization under parametric choice models. The section ends with a brief introduction of nonparametric approaches to assortment optimization.

Kök et al. (2008) provide an extensive survey on both assortment optimization and inventory planning. They start their survey by briefly reviewing four streams of literature that assortment planning models build on, namely product variety and product line design, shelf space allocation, multi-product inventory systems, and a consumer's perception of variety. Next, the authors discuss consumer substitution behaviour and introduce three popular demand models—multinomial logit, exogenous demand, and locational choice. They present selected literature on assortment planning related to the basic problem, as well as extensions thereof including supply chain considerations, demand learning, and assortment changes during the selling season or multi-category assortment planning. Finally, the authors discuss demand and substitution estimation methodologies, present industry approaches to assortment planning of four retailers, and compare these industry approaches with academic ones.

Hübner and Kuhn (2012) provide an excellent review on integrated assortment and shelf space planning in retail category management. They classify the literature on assortment problems by the underlying demand model, substitution reasons (e.g. out-of-assortment, out-of-stock), solution method, and maximum number of items considered in the test case and provide additional information regarding model enhancements considered in the listed studies.

Later on, Karampatsa et al. (2017) conduct a survey on assortment and shelf-space planning models in retail category management. The literature on assortment problems is mainly classified according to the objective (assortment, inventory, price), the underlying demand model, the type of substitution, and the solution method that has been used in each model, along with the average number of items used in the test cases. The considered literature on assortment problems is limited to multinomial logit, exogenous demand and locational choice demand models.

Finally, Berbeglia et al. (2021a) review selected choice models and provide an application guideline for them by defining suitable operational environments for each of the considered choice models based on extensive numerical studies. They

extend their studies by empirically evaluating the revenue performance of the considered choice models.

In contrast with all above-introduced surveys, we aim at reviewing literature on assortment optimization problems in general, without focusing on a particular discipline such as revenue management or retail operations. Moreover, in contrast with Hübner and Kuhn (2012) and Karampatsa et al. (2017), we focus on classic, pure assortment problems and do not consider extensions such as shelf-space planning. In addition, we do not limit our review to selected choice models. Finally, all of the reviews mentioned before provide their review of the existing literature in a textual form describing individual articles' contributions to the topic. To the best of our knowledge, only the two reviews by Hübner and Kuhn (2012) and Karampatsa et al. (2017)—both covering the topic of assortment and shelf-space planning—additionally provide a small tabular classification of selected literature.[2] Instead of describing individual articles' contributions, our systematic literature review should provide a structured overview of assortment optimization settings available in the literature by classifying existing articles according to a proposed taxonomy covering a broad range of factors related to assortment optimization. This makes it easy for academics and practitioners alike to determine the assortment optimization setting that is most suitable for them and identify relevant related literature. Hence, the aim is not to provide a detailed discussion of all existing approaches to assortment optimization or their underlying choice models, but rather to give an overview of different studied settings and to identify existing research gaps.

The remainder of this review is structured as follows: Our proposed taxonomy to group the available literature on assortment optimization concerns factors related to the optimization problem itself, the modelled customer behaviour, and the solution concept applied for solving the problem at hand. Each of these topics is comprehensively addressed in Sects. 2, 3, and 4, respectively. In Sect. 5, we describe the procedure of conducting our systematic literature review, summarize our proposed taxonomy, assemble an overview of the literature on assortment optimization, and analyse it based on a selection of key factors. We provide future research directions in Section 6 and conclude our review in Section 7.

2 Modelling assortment optimization problems

This section is targeted to the introduction of different versions of the assortment problem. In Sect. 2.1, we start by introducing the classic assortment problem, followed by the presentation of robust and dynamic versions thereof in Sects. 2.2 and 2.3, respectively. We terminate the section by a concise description of a variety of constraints that are typically considered in the assortment optimization literature in Sect. 2.4.

[2] To be precise, both reviews consider the factors demand model, model enhancements, substitution reason (out-of-assortment/out-of-stock), solution method, and number of items in case study, whereby Karampatsa et al. (2017) additionally report the model consideration (Assortment, Inventory, Pricing).

2.1 Classic assortment problem

Assortment optimization refers to the problem of determining a selection of options to be offered to arriving customers in order to maximize a given objective, typically the expected revenue (Mišić and Perakis 2019).[3] This is sometimes also referred to as assortment planning. In case the assortment is specifically tailored to individual customers, the problem is called assortment personalization (see e.g. Golrezaei et al. 2014). More formally, let $N = \{1, ..., n\}$ be the set of available items and denote the no-purchase option by $\{0\}$. Then, $N \cup \{0\}$ refers to the selection of available items including the no-purchase option. The retailer needs to select a subset of the available items to be offered. Following Davis et al. (2013), instead of denoting the offer set as an actual subset, the assortment can be represented by a binary decision variable x_i for each item $i \in N$ that indicates whether this item is offered or not by setting $x_i = 1$ and $x_i = 0$, respectively. More formally, define

$$x_i = \begin{cases} 1, \text{if item } i \text{ is offered} \\ 0, \text{else} \end{cases} \qquad \forall i \in N.$$

Note that the no-purchase option is always offered, implying that $x_0 = 1$ holds for any assortment. The number of options that are offered in an assortment can be obtained by summing over all x_i, i.e. $\sum_{i \in N} x_i$.

The demand for any option depends on the preferences of the customers and is captured by a choice model specifying the probability that a customer selects a particular option from a given offer set as detailed in Sect. 3. Assuming a general choice model—i.e. any arbitrary choice model such as the multinomial logit, the nested logit, the Markov chain choice, or the rank-based model without specifying a particular one (Talluri and van Ryzin 2004; Gallego and Topaloglu 2019; Strauss et al. 2018)—the customer selects alternative $i \in N$ with probability $p_i(\boldsymbol{x})$ given that assortment $\boldsymbol{x}$ is offered and decides to not purchase anything with probability $p_0(\boldsymbol{x})$.

For all offer sets $\boldsymbol{x} \in \{0, 1\}^n$, the purchase probabilities need to satisfy certain requirements (see Talluri and van Ryzin 2004). First, the choice probabilities of all items $i \in N$ and the no-purchase option need to be non-negative. Second, the choice probabilities of all items and the no-purchase option must sum up to one and third; the choice probability of an item must equal zero if this option is not contained in the offer set. These requirements can be formalized as follows:

1. $p_i(\boldsymbol{x}) \geq 0 \; \forall i \in N \cup \{0\}$
2. $p_0(\boldsymbol{x}) + \sum_{i \in N} p_i(\boldsymbol{x}) = 1$
3. $p_i(\boldsymbol{x}) = 0$ if $x_i = 0$

[3] For simplicity, we choose the expected revenue as running example within this review. However, note that all concepts can be analogously applied for different objectives.

Moreover, assume that item $i \in N$ is sold at revenue r_i and that the market is of size 1 without loss of generality. Then, the classic assortment problem targeted at optimizing the expected revenue is denoted by

$$\begin{array}{ll} \max_{\boldsymbol{x}} & \sum_{i \in N} p_i(\boldsymbol{x}) \cdot r_i \\ \text{subject to} & x_i \in \{0,1\} \qquad \forall i \in N \end{array} \tag{AOP}$$

In the above optimization problem (AOP), $\boldsymbol{x} \in \{0,1\}^n$ denotes the decision variable indicating which of the items $i \in N$ are included in the offered assortment. The objective represents the expected revenue and sums up the expected revenue obtained by each item i. The expected revenue per item is obtained by multiplying the purchase probability $p_i(\boldsymbol{x})$ of the item with its revenue r_i. The optimization problem formulation is completed by a binary constraint ensuring that all values of x_i are either 0 or 1.

Note that such assortment optimization problems are combinatorial by their nature. The number of possible assortments to be evaluated is

$$\binom{n}{1} + \binom{n}{2} + \cdots + \binom{n}{n}, \tag{1}$$

where $\binom{n}{k} = \frac{n!}{k!(n-k)!}$ for any integer $1 \leq k \leq n$. Since the possible number of combinations quickly explodes, solving the problem by full enumeration is hardly possible. However, the above way of formulating the optimization problem allows to implement and solve it using standard solvers such as CPLEX or Gurobi.

2.2 Robust assortment problem

Assortment optimization problems are typically based on an underlying probabilistic choice model. The parameters of the choice model are mostly assumed to be unknown and thus need to be estimated from data, such that statistical errors in the parameters are unavoidable (Désir et al. 2023). The optimal assortment decision is then made based on the estimated parameter values while ignoring any uncertainty associated with these estimates (Rusmevichientong and Topaloglu 2012).

To overcome this issue, an uncertainty set, i.e. a set of likely parameter values, can be considered, which includes the true parameters with high confidence based on the statistical estimation procedure. Then, the overall goal is to determine a revenue maximizing assortment while explicitly accounting for the uncertainty in the choice model parameters. Instead of directly optimizing the expected revenue as done in (AOP), this can be achieved by maximizing the worst-case expected revenue, where the worst case is taken over all possible parameter values in the uncertainty set. The resulting optimization problem is referred to as robust assortment optimization and can be formalized by

$$\begin{array}{ll} \max_{\boldsymbol{x}} \min_{\boldsymbol{\Theta}} & \sum_{i \in N} p_i(\boldsymbol{x}, \boldsymbol{\Theta}) \cdot r_i \\ \text{subject to} & x_i \in \{0,1\} \qquad \forall i \in N \end{array} \tag{Robust AOP}$$

where $\boldsymbol{\Theta}$ represents the model parameters of the underlying choice model (Rusmevichientong and Topaloglu 2012).

Note that we made the dependence of the choice probabilities on the selected model parameters explicit in the objective of the above problem formulation (Robust AOP). Moreover, note that (Robust AOP) maximizes the minimum expected revenue, where the decision variables for the minimization are the model parameters; the decision variable of the subsequent maximization of the worst-case expected revenue is again the binary assortment vector $\boldsymbol{x} \in \{0,1\}^n$. As before, the optimization problem formulation includes a binary constraint ensuring that all values of x_i are indeed either 0 or 1.

A popular way of approaching robust assortment problems is to make use of duality results. For example, Li and Ke (2019) and Mehrani and Sefair (2022) utilize strong duality for constructing their solution methods to robust assortment problems under the multinomial logit and the ranking-based choice model, respectively. Likewise, Désir et al. (2023) study the robust assortment problem under the Markov chain choice model and propose an iterative algorithm that makes use of the min-max duality.

2.3 Dynamic assortment problem

The classic static assortment optimization introduced in Model (AOP) assumes that the customers' preferences are known or can be estimated from data and do not change over time. In this case, the assortment problem is targeted to determine a revenue-maximizing assortment that is offered over the whole selling season. In contrast, in dynamic assortment optimization the customers' choice behaviour is unknown a priori and must be learned step by step by sequentially offering different trial assortments to arriving customers over a certain time horizon and observing the corresponding click or purchase behaviour. For simplicity, it is typically assumed that exactly one customer arrives per selling period $t = 1, ..., T$ and is offered a period-specific assortment $\boldsymbol{x^t}$. This setting is e.g. relevant when the seller follows a multi-period planning horizon or for short-lived items without sufficient historical data, see e.g. Caro et al. (2014).

It is realistic that the assumed customer behaviour and thus the offered assortment are incorrect in the beginning of the selling season and improve over time by observing more and more customer behaviour. However, this so-called exploration period should not be too long as the offered trial assortments might be suboptimal and thus lead to lower revenues. Therefore, the decision maker at each time step faces the decision whether to keep exploring more assortments to better learn the customer behaviour or begin to exploit the best assortment determined so far. Clearly, the longer the exploration period, the greater the chance to find a near-optimal assortment. But this long exploration period might come with large accumulated regret, i.e. large cumulative expected revenue losses caused by offering suboptimal assortments. Hence, the

question is how much time should be spent in learning customer preferences before exploiting the best assortment determined till then. This problem is referred to as exploration–exploitation trade-off and is characteristic for this type of dynamic assortment problem, see Caro and Gallien (2007). The overall goal in such dynamic settings is to minimize the cumulative regret or to maximize the expected cumulative revenue over the whole selling horizon, see e.g. Rusmevichientong et al. (2010) and Bernstein et al. (2019).

Besides this multi-period problem formulation, there exist further assortment optimization specifications that unfold dynamically. In classic assortment optimization settings, the whole assortment is simultaneously offered to arriving customers with the goal of maximizing the expected revenue. However, there exist numerous settings where it might be overwhelming to present a customer with a large number of possible options all at once. This is e.g. the case in appointment scheduling when booking doctor's appointments or a table in a restaurant. In these cases, it can be beneficial to provide the customer with only a handful of time slots in consecutive stages until the customer identifies a suitable time slot. Besides this, e-tailers also often make use of sequential offerings, particularly for product recommendations or when displaying search results across multiple results pages, see e.g. Liu et al. (2020).

In such settings, the purchase dynamics of a customer unfold sequentially over T stages. In each stage $t = 1, ..., T$, one assortment $\boldsymbol{x^t}$ of items is selected and made available for purchase. Moving from one stage to the next, the customer either decides to purchase one of the items offered in the present stage according to his choice model preferences and leave the system or to not make a purchase at that time. In the former case, the seller gains an option-specific revenue; in the latter case, the customer can progress to the next stage if any is left or leave the system without making a purchase. The purchase decision can be either governed by a stage-dependent choice model reflecting the fact that customers' preferences could change from stage to stage due to e.g. updated perceptions or patience waning, or by a stage-invariant choice model that is used across all stages (Liu et al. 2020; Feldman and Segev 2022). Many authors propose adjusted versions of known choice models to capture this dynamic behaviour (see e.g. Feng and Wang 2021; Flores et al. 2019).

Overall, the dynamic version of the assortment problem can be formalized by

$$\begin{array}{ll} \max_X & \sum_{t=1}^{T} \sum_{i \in N} p_i^t(\boldsymbol{x^t}) \cdot r_{it} \\ \text{subject to} & x_i^t \in \{0,1\} \qquad\qquad \forall i \in N, t = 1, ..., T \end{array} \tag{Dynamic AOP}$$

where x_i^t indicates whether option i is offered in period/stage t, p_i^t denotes the purchase probability of option i in period/stage t, and r_{it} represents the revenue for option i in period/stage t. Note that the objective of the above problem formulation (Dynamic AOP) entails a double sum. In the inner sum, the expected revenue is calculated for each t separately. Subsequently, the outer sum determines the total cumulative expected revenue by summing the expected revenues across all considered sales periods/stages $t = 1, ..., T$. The decision variable of this optimization problem is the binary assortment matrix $X \in \{0,1\}^{T \times n}$ with rows $\boldsymbol{x^t}, t = 1, ..., T$. Finally, note

that the above optimization problem formulation (Dynamic AOP) includes a binary constraint ensuring that all values of x_i^t are indeed either 0 or 1.

2.4 Constraints

In practice, assortment tasks are often accompanied by a broad set of requirements. These requirements can be incorporated in the optimization problem as constraints, i.e. as logical conditions to be satisfied by the solution of the optimization problem at hand. In assortment optimization, typically hard constraints are considered instead of soft ones. The former puts conditions on the variables that must be satisfied, whereas violating the latter merely imposes a penalty on the cost function. There exists a variety of constraints that are considered in the literature on assortment optimization. They are briefly introduced in the following.

- The cardinality constraint limits the total size of the offered assortment to a maximum of C options; more formally $\sum_{i \in N} x_i \leq C$. To avoid trivial cases, typically $C \leq n$ is assumed. Assortment optimization under such a cardinality constraint is e.g. studied by Lo and Topaloglu (2021).
- Under a capacity constraint—sometimes also referred to as space constraint or knapsack constraint—each option i is associated with an item-specific weight or size w_i and the capacity constraint limits the total available weight/space to C; more formally $\sum_{i \in N} w_i x_i \leq C$. Such a capacity constraint is e.g. considered in Feldman and Topaloglu (2017b). Note that for the special case of uniform weights $w_i = 1 \; \forall i$, the capacity constraint reduces to a cardinality constraint.
- Totally unimodular (TU) constraints refer to certain types of constraints whose combination results in a constraint system that exhibits the so-called total unimodularity property. The constraint system is of the form $A\boldsymbol{x} \leq \boldsymbol{b}$ where A satisfies the TU property—i.e. A is a matrix with every square submatrix having determinant ± 1 or 0—and $\boldsymbol{b}$ denotes a vector that is assumed to be integral. This constraint structure subsumes a variety of different constraints such as the described cardinality and capacity but also precedence and partition constraints (see Davis et al. 2013).
  - Precedence constraints assume that a product i can only be offered to customers in case a certain other product j is also offered. This constraint can be formalized as $x_i \leq x_j$.
  - Under partition constraints, the products are partitioned into K disjoint groups $S_1, ..., S_K$ and there is a limit b_k on the number of products offered per group k, i.e. $\sum_{i \in S_k} x_i \leq b_k$.
- Inventory constraints limit the number of units of a product that can be sold. This constraint is particularly important in the context of (Dynamic AOP), which comprises the determination of revenue-maximizing assortments over a whole selling season of length T. Dynamic assortment optimization under such an inventory constraint is e.g. studied by Rusmevichientong et al. (2020).

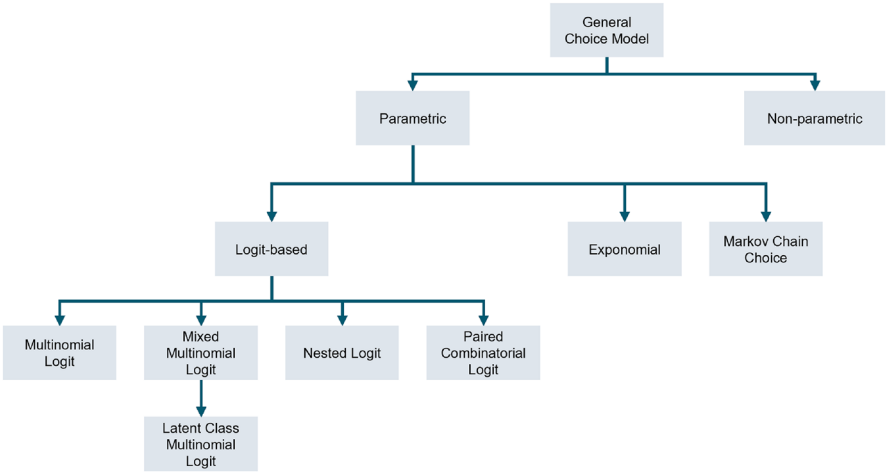

Fig. 2 Tree structure of choice models

3 Capturing customer behaviour

This section is targeted to the introduction of different aspects related to customer behaviour. In Sect. 3.1, we start by describing the most popular choice models and frequently used estimation techniques. Subsequently, model extensions related to the allowed number of item purchases or the incorporation of consideration sets are discussed in Sects. 3.2 and 3.3, respectively. We terminate the section by a description of the impact of different sales channels on customer behaviour in Sect. 3.4.

3.1 Choice model design and estimation

As mentioned before, assortment optimization refers to the problem of determining an optimal assortment of options that should be offered to the customers in order to maximize the expected revenue with respect to a given choice model (Mišić and Perakis 2019), that is to say an assumption on the customers choice preferences. Intuitively, choice models are used to capture the demand behaviour of the customers and thus can be used to model which of the offered options might be purchased by the customers. Doing so, the optimal assortment needs to find a balance between including options and cannibalizing the demand of other options' sales, see Kök et al. (2008).

In Sect. 2, we introduced the assortment optimization problem under a general choice model that provides the choice probabilities for all products given a certain assortment. However, to solve an optimization problem, a concrete choice model must be chosen. In recent years, the assortment problem has been studied under a variety of choice models.

As depicted in Fig. 2, choice models can be divided into parametric and non-parametric approaches whereby the parametric choice models can be subdivided into logit-based choice models such as the multinomial logit, the mixed multinomial

logit, the nested logit, and the paired combinatorial logit model and further parametric choice models including the exponomial and the Markov chain choice model (see Strauss et al. 2018). Both, the parametric and the nonparametric approaches are briefly presented in Sects. 3.1.1 and 3.1.2, respectively. A more detailed introduction of these choice models is found in Appendix A.

3.1.1 Parametric choice models

Parametric choice models are fully defined by a finite number of parameters that do not scale with the number of offered items (Berbeglia et al. 2021a). Such models are typically based on random utility theory, where it is assumed that consumers associate a certain utility with every item, and decide on the alternative that maximizes their utility (Strauss et al. 2018). This framework is also referred to as random utility maximization (RUM). The utility of an option is assumed to be composed of a deterministic and a random component. Different assumptions made on the distribution of the random component result in different choice models. Below, we briefly present a selection of the most common parametric choice models considered in the literature on assortment optimization. A more detailed introduction of the RUM framework and the below presented choice models can be found in the textbooks of Ben-Akiva and Lerman (1985), Train (2009), and Hensher et al. (2005) as well as in Appendix A.

- **Multinomial logit (MNL):** The multinomial logit model of Luce (2012) and McFadden (1973) can be used to estimate the probabilities of different possible choice options of a customer based on a selection of given attributes. As an example, imagine a customer can decide to purchase a t-shirt either made of cotton or silk or neither of them. Based on price and quality, the customer associates a preference weight with each of these products and the probability to select an option is determined by this options preference weight relative to the total preference weight of the offer set. Nevertheless, it should be taken into account that the MNL model might have a deficiency in representing the choice among alternatives with shared attributes—the Independence of Irrelevant Alternatives (IIA) property (see Ben-Akiva and Lerman 1985) illustrated by the well-known 'red bus/blue bus' paradox (Debreu 1960)—and should therefore be used with caution according to Talluri and van Ryzin (2004).
- **Mixed multinomial logit (MMNL):** The mixed multinomial logit choice model (McFadden and Train 2000) considers different customer segments whereby the preferences of each segment follow a segment-specific MNL model. Imagine for example two customer segments—budget conscious shoppers and quality-focused consumers. Customers within the former segment put more weight on the product price, whereas customers from the latter segment are less focused on the product price but more on its quality. Such MMNL models are able to approximate the choice probabilities of any choice model within the RUM framework arbitrarily close under mild regularity conditions (McFadden and Train 2000). The latent class multinomial logit (LC-MNL) is a special case of the MMNL where the random MNL parameters follow a discrete distribution.

- **Nested logit (NL):** Under the nested logit model, it is assumed that the choice set can be partitioned into disjoint subsets called nests (Heiss 2002) in a way such that the IIA property holds within each nest but not across different nests (Strauss et al. 2018). Then, the choice probability for a certain option is the product of the probability to choose some alternative from the same nest in which this certain option is located and the conditional probability to choose exactly this certain option given some alternative in the same nest as this certain option is selected. Imagine for example a customer who is looking for new clothes, which can be separated into the categories business wear and casual clothing. The customer decides to purchase a black suit belonging to the business wear category.
- **Paired combinatorial logit (PCL):** Under the PCL model, all items are grouped into nests of size two, whereby the model allows for correlations between the utilities of any pair of items and is thus able to capture situations where the preference of a customer for a particular item offers insights into the customer's attitude towards another item. Under this model, the probability that a certain option is chosen is obtained by summing over all nests of size two that contain this option. To be precise, one sums the product of (i) the probability that a customer picks the nest of size two and (ii) the probability that the certain option is selected given that an alternative from the nest of size two is purchased (see Koppelman and Wen 2000). As an example, imagine a customer is interested in a blue t-shirt. This customer likely shows an affinity towards blue clothes and thus might be interested in blue shorts as well.
- **Exponomial (EXP):** Under the exponomial choice model proposed by Alptekinoğlu and Semple (2016), the choice probabilities are expressed as a linear combination of exponential terms—hence the name 'exponomial' (Strauss et al. 2018). In contrast with the MNL or the NL model where the customers' willingness to pay distribution is assumed to be positively skewed, the EXP model assumes a negatively skewed distribution of customer utilities. This model is particularly suitable for situations in which the customer is well informed about products and their values such that his willingness to pay distribution is negatively skewed because he would be deterred by the prospect of overpaying (Alptekinoğlu and Semple 2016). Imagine for example a person buying a new, expensive wristwatch. This customer likely knows the watch's MSRP as well as further offer prices across different online sales platforms. Hence, the customer obtains a benchmark price for all watches in his choice set and the likelihood that he is willing to overpay this benchmark price is way lower than the likelihood that he is willing to underpay his benchmark price, suggesting a negatively skewed willingness to pay distribution.
- **Markov chain choice (MCC):** The Markov chain choice model proposed by Blanchet et al. (2016) can approximate any discrete choice model within the RUM framework under mild assumptions (see Blanchet et al. 2016). The model represents the customer choice process by a Markov chain where each state corresponds to a product or the no-purchase option. Every product state is connected with state 0 representing the no-purchase option. The customer arrives at a state according to its arrival probability. When arriving at a certain state, the customer purchases the corresponding product in case it is offered. Otherwise, the cus-

tomer proceeds to another state with a certain transition probability. Such a transition probability can be thought of as the probability to substitute one product with the other in case the former is unavailable (see Strauss et al. 2018).

3.1.2 Nonparametric choice models

The previously introduced parametric choice models fully depend on the choice of their underlying parameters. These parameters are typically unknown and need to be chosen or estimated in practice. Likewise, the attributes driving the choice process need to be selected, which is a potential source of specification errors (Strauss et al. 2018). Moreover, parametric choice models assume that the choice behaviour can be captured by a given functional form. Yet, the specified functional form may not adequately capture the actual choice behaviour (see Strauss et al. 2018).

Nonparametric choice models by design do not suffer from these problems as they are not built upon any assumption on the data structure but are solely shaped by data. However, nonparametric choice models typically do not allow for extrapolation and prediction of changes in the demand pattern due to changes in an options attributes (Berbeglia et al. 2021a).

Such nonparametric models are typically designed as ranked lists of preferences, also referred to as customer types. Under rank list-based models, the customer chooses the highest ranking available item or leaves without purchase if none of the offered items ranks higher than the no-purchase option. Demand is then modelled by a probability distribution over all customer types. Overall, this model is quite general and subsumes various choice models typically considered in assortment optimization such as the MNL (see Mahajan and van Ryzin 2001).

3.1.3 Estimation

As mentioned before, the parameters of the parametric choice models introduced so far are typically unknown and therefore need to be chosen. The same holds for the descriptors of nonparametric models such as the empirical distribution of a demand function. For this purpose, data are required, whereby the data used for such estimation tasks can consist of stated-preference or revealed-preference data. Stated-preference data comprise data that are based on behavioural intentions and responses to hypothetical choice situations, whereas revealed-preference data describe actual customer behaviour (see Ben-Akiva 1994). In the area of assortment optimization, typically historical sales data reflecting actual customer behaviour—i.e. revealed-preference data—are used for this purpose. However, sometimes it might be the case that no or not sufficient historical data are available. Then, stated choice experiments can be used to obtain stated-preference data on the customers behaviour within hypothetical choice situations.

In literature and practice, there exist two approaches that are particularly popular for performing estimation tasks—maximum likelihood estimation (MLE) and expectation–maximization (EM).

- **Maximum likelihood estimation** is a method used to estimate the unknown parameters by maximizing a likelihood function such that the observed data are most probable under the assumed model (see e.g. Hensher et al. 2005). Its solvability in closed form is only given in certain special cases. A well-known alternative to MLE is the so-called least squares minimization. Berbeglia et al. (2021a) provide empirical evidence that MLE and least squares minimization have comparable performance in terms of out-of-sample prediction accuracy for all considered choice models though MLE tends to be slightly superior in the majority of all analysed scenarios.
- **Expectation–maximization**—proposed by Dempster et al. (1977)—is an iterative method to determine maximum likelihood estimates of unknown parameters in statistical models by alternating between expectation (E) and maximization (M) steps. In the former step, a function for the expectation of the log-likelihood is created and evaluated using the current parameter estimates. In the M-step, the parameters maximizing the expected log-likelihood function found in the E-step are computed to obtain improved parameter estimates.

In recent years, the rise of machine learning has also affected the area of choice modelling. According to van Cranenburgh et al. (2022), machine learning advanced considerably when it comes to estimation algorithms that are able to deal with large volumes of data and complex model specifications. These algorithms can also be employed for the estimation of choice models. For example, Lederrey et al. (2021) propose new efficient stochastic optimization algorithms that are able to deal with large data sets to estimate discrete choice models.

Another common approach of combining choice modelling and machine learning is based on a two-step procedure. First, the utility is modelled as a function of (product/customer) features using a machine learning method, and second, the utilities are related to the choice probabilities using a discrete choice model such as the MNL or the NL (see e.g. Cai et al. 2022). For example, Han et al. (2022) and Sifringer et al. (2020) both propose to replace the utility by a neural network function of product (and customer) features. The utilities are then mapped to the choice probabilities via a MNL model in the former study and via both, MNL and NL models in the latter one.

Likewise, Doudchenko and Drynkin (2020) also propose a two-step procedure. In their case, first a prediction problem linking the individual level variables to choice probabilities is solved. Second, a standard discrete choice model is estimated to find coefficients at pre-defined variable values, which enables the gain of the coefficients for any other point by solving a system of linear equations. For more information regarding the intersection of choice modelling and machine learning, we refer the interested reader to van Cranenburgh et al. (2022) and the references therein as well as to Sect. 6, where we comment upon future research on combining demand modelling, machine learning and assortment optimization.

3.1.4 Empirical performance

The selection of a choice model that is suitable for the given operational context in terms of model specification, computational tractability, and prediction accuracy is challenging. In their empirical study, Berbeglia et al. (2021a) analyse nine choice models extensively used in the assortment optimization literature, namely the multinomial logit, the mixed logit, the latent-class multinomial logit, the nested logit, the exponomial, the Markov chain choice model including all possible transactions, a reduced Markov chain choice model where the transitions are designed following a vertical differentiation of the items, a Markov chain choice model with transition matrix of rank two, and a rank list-based choice model. These models are compared with regard to their predictive ability and the computational time required to estimate different models.

- Regarding the prediction accuracy, the authors find that the exponomial model stands out in small training data environments, whereas in large training data environments, the Markov chain choice model by far exhibits the best performance. This observation holds for all three types of evaluated instances—synthetic, semi-synthetic, and real. Moreover, according to the study the Markov chain model consistently appears among the top three performers when data volume increases and profits the most from increasing data volumes. In addition, the authors find that all models except for the rank list-based model improve their predictive performance when the consistency of the customer preferences is low, i.e. when many different customer types exist. In contrast, the predictive performance of all considered choice models deteriorates with larger assortments as this setting provides less substitution patterns since a big fraction of consumers get their most preferred option.
- Regarding the computation time, the study of Berbeglia et al. (2021a) provides evidence that the MNL is by far the fastest choice model to estimate. Nested logit and exponomial choice models are on average ten times slower to estimate than the MNL model. All other models considered in this study, namely Markov chain, rank-list, latent-class MNL, and mixed logit, are on average at least 100 times slower to estimate compared to the MNL though the authors expect this gap to increase when larger data sets are used.

3.2 Number of item purchases

The previously introduced choice models characterize different customer behaviour, though the customer behaviour is not only captured by the selection of a choice model type but also by the decision of the customer how many products should be purchased within a single visit.

Most studies focus on the single-purchase case. In this setting, each customer is assumed to buy at most one product. In practice, this is e.g. the case when purchasing luxury goods such as a vehicle. Though, the single-purchase setting also covers the case when multiple copies of the same product are purchased.

However, recently, there is an increasing amount of researchers focusing on the multi-purchase version of the assortment problem where the customer not only decides whether and which product to buy but also on the number of different products to be bought. Doing so, researchers typically relax the model assumption that customers only choose at most one product per visit from the offer set and instead propose multi-purchase choice models allowing customers to purchase more than one product at a time, see e.g. Bai et al. (2023a) and Tulabandhula et al. (2023).

Note that in practice there are plenty of scenarios where customers purchase multiple products at a time. Imagine for example a customer shopping clothes or accessories. This customer often buys multiple pieces within the same visit. Another example are online shops with delivery costs that are waived if a certain purchase price threshold is met. In these cases, customers often purchase multiple products to get rid of the delivery costs. When purchasing several products at a time, customers can either select multiple versions or copies of the same item or buy completely different items. According to Bai et al. (2023a) who analyse sales data of a leading flash-sales e-retailer, over 89% of the customers purchasing two products indeed purchase two different products. For these multi-purchase scenarios, choice models that are based on the single-purchase assumption might not perform well, see Feldman et al. (2021) and Wang et al. (2023d).

3.3 Consideration sets

Customer demand is typically estimated using choice models that rely on information regarding what customers do and do not purchase. To calibrate such demand models, e.g. sales transaction data in case of retail operations and revenue management or bookings from past interactions between peers in case of online platforms are used. Given these data, classic choice models are trained based on the assumption that the chosen option is preferred over all other items in the offer set, see Jagabathula et al. (2023).

However, the decision of a customer to not purchase an item must not necessarily result from the fact that this item is not offered but can also result from the fact that it is not considered (Jagabathula et al. 2023). In practice, it is well known that customers do not directly choose from the whole available assortment. Instead, they use a set of simple rules to first quickly shrink the set of offered items to a small subset of options that are most interesting for them and then choose from this small subset of remaining options, which is referred to as consideration set (see e.g. Aouad et al. 2020). For instance in retail, imagine a customer selecting from the jacket category. This customer may not evaluate the full offered jacket assortment but only consider a subset of jackets in the desired size that are priced within the affordable budget.

Models ignoring such consideration sets assume that the chosen option is preferred over items that are not even considered, which might lead to model bias (Jagabathula et al. 2023). Existing literature proposes so-called consider-then-choose (CTC) models to overcome this issue and account for the behaviour of first setting up a consideration set and subsequently choosing an item from the intersection of offer set and consideration set (see e.g. Aouad et al. 2020). Such consider-then-choose

approaches originate from empirical literature in marketing and psychology. To be precise, the idea of whittling down choices into consideration sets is first proposed by Campbell (1969) and formulated into a theory on customer behaviour by Howard and Sheth (1969). The incorporation of consumers' consideration sets can improve both the explanatory and the predictive power of demand models, which in turn helps to enhance assortment decisions. The literature on assortment optimization under a consider-then-choose model comprises a broad range of different consideration set structures, see e.g. Aouad et al. (2020) and Jin et al. (2023).

Despite their intuitive behaviour, CTC models are difficult to estimate in practice as one typically only knows the offer set and the customers' choice. A customers' consideration set is mostly not observable in practice and could be any subset of the full offer set or the category containing the chosen option. Hence, common choice models often assume that offer set and consideration set are equivalent, see Jagabathula et al. (2023). However, Jagabathula et al. (2023) analyse CTC models using both synthetic and real-world data sets and find that CTC models outperform classic choice models in cases when the offer set is not perfectly observable.

3.4 Sales channel

Assortment optimization finds important application in both online and offline channel settings, though existing work on assortment optimization mostly provides guidance on how firms should optimize their offerings in single-channel settings. Online settings involve online sales of products or services, whereas offline sales take place in physical stores or outlets. Those two settings differ in terms of both, customer experience and data and modelling topics. We briefly discuss both factors in the following, starting with the customer experience.

- **Assortment size:** Online channels are typically able to provide the customer with larger assortments compared to offline channels. The assortment size influences the substitution behaviour of customers as e.g. larger assortments imply less substitution since customers are more likely to find their most preferred options anyway. However, larger assortments often result in higher search efforts till the desired item is found, which might decrease the items utility. To account for search efforts, choice models in online settings often incorporate search costs. Moreover, online settings often apply cardinality constraints to limit the assortment size for keeping customers attention and reducing search efforts.
- **Opening hours:** Online channels are mostly available 24 h a day without closing such that customers can shop at any time. This allows for spontaneous purchases but might also lengthen the purchase decision as customers can easily decide to quit and return later for purchase. In contrast, offline stores only offer limited opening hours.
- **Delivery:** In contrast with offline channels where the customer can typically directly take the purchased option, online retail stipulates that the purchased option needs to be delivered. This might lead to issues during order fulfillment

such as wrong or broken delivery or too long delivery times which in turn might lead to customer dissatisfaction.

- **Product presentation:** In offline channels, options are exhibited on shelves where customers are able to grasp all available options at a time. In contrast, in online settings the available options are often offered sequentially across multiple results pages. The former case is modelled by classic assortment optimization approaches, whereas the latter one requires a sequential, dynamic approach.
- **Product interaction:** In contrast with offline channels where the customer is able to physically check the desired products, online channels come with an increased amount of uncertainty as customers are not able to directly experience the product. Imagine a customer wants to buy a t-shirt. In online shopping, this customer is not able to look at, touch, and try on the t-shirt but needs to rely on virtual experience based on pictures or short video clips. Consequently, there is an increased uncertainty in terms of e.g. size or colour involved in online shopping. This uncertainty might be intensified by other factors such as product misspecification, misrepresentation, and misleading advertisement and can e.g. be incorporated via the random component of RUM-based choice models.

Besides the customer experience, online and offline channel also differ from a data and modelling perspective as briefly expounded in the following.

- **Data availability:** Online channels profit from increased availability of customer related information and data. The former is typically obtained from customer profiles comprising information such as age, gender, and location. The latter mostly consists of historic click and purchase behaviour of the customer. The availability of such personal data allows for personalization in online channels. That is to say the seller can dynamically adjust the sales strategy for individual customers by immediately providing the customer with a selection of relevant options upon website arrival. Doing so, the customer benefits from being offered an assortment of suitable options and the seller profits from increased sales due to personalized assortments.
- **Modelling challenges:** The sheer mass and dimensionality of data available in online settings come with modelling challenges as demand models are not necessarily able to deal with large amounts of data, i.e. with a high number of observations, or high-dimensional data, i.e. data containing lots of different information for each customer. However, research is recently devoting increased attention towards this area such that first approaches for dealing with high-dimensional data are already available (see e.g. Miao and Chao 2022; Wang et al. 2023c; Kallus and Udell 2020). In addition, online channels suffer from the exploration-exploitation dilemma. On the one hand, the seller aims at conducting as much exploration as possible to learn the choice models by offering diversified assortments. On the other hand, extensive experimentation could harm exploitation in terms of maximizing revenues.

The advantages of both online and offline channel can be combined by considering an omni-channel setting. In this case, the firm is able to offer a wide range of

options via their online channel and additionally allows customers to experience the touch and feel of product attributes in offline stores before purchase. In this setting, the selection of options offered via the offline channel impacts the online purchase behaviour.

4 Solving assortment problems

One typically distinguishes different types of solution concepts that we briefly introduce in Sect. 4.1. Moreover, since research recently focuses on approximation-based approaches that provide performance guarantees, we detail on the latter in Sect. 4.2 where we additionally comment on the empirical evaluation of such performance guarantees.

4.1 Solution concepts

Existing solution concepts can be divided in two groups: exact optimization methods and non-exact optimization methods, whereby the latter group can be further split into heuristics and approximation algorithms. All of those concepts are briefly discussed in the following:

- **Exact optimization methods** guarantee to find an optimal solution. In the literature and practice, there exist various different solution approaches for determining the exact solution of an optimization problem. Two of the most popular ones are full enumeration and the usage of standard solvers.
  - **Full enumeration** refers to the approach when all possible assortments are enumerated and evaluated in terms of their revenue performance. Clearly, the optimal assortment can be found by selecting an assortment that yields the best performance. However, the formula for determining the number of possible assortments provided in Eq. (1) shows that $n = 10$ items already yield 1023 possible assortments; $n = 100$ items even result in $1.2676506002282297 \cdot 10^{30}$ possible offer sets. Obviously, the problem quickly explodes such that complete enumeration of all possible combinations becomes intractable—even when the solution space is reduced by applying selected constraints. Still, for practical applications with a small number of items, full enumeration may represent a reasonable approach, because it can be combined even with complex choice models using simulation or neural networks.
  - **Standard solvers** like Gurobi or CPLEX can be used to solve e.g. linear, quadratic, mixed-integer, or quadratic-constrained programs. Many assortment optimization problems can be (re)formulated in one of these ways and are thus solvable using such standard solvers. For example, Davis et al. (2013) show how to transform an (AOP) under the MNL model and TU constraints into a linear program. Likewise, Haase and Müller (2014) discuss three linear reformulations of originally nonlinear facility location

problem formulations under the MNL model in terms of solvability. A survey on corresponding techniques is also given in Bechler et al. (2021). Furthermore, standard solvers can be applied in the context of approximation algorithms based on rounding techniques.

- **Non-exact optimization methods** do not necessarily yield an optimal solution. This group of solution concepts captures heuristics and approximation-based methods, both of which are briefly introduced in the following:
  - **Heuristics** are optimization methods that try to provide a good but not necessarily optimal solution. In operations research, there has been an enormous study of various types of heuristics, including construction and improvement heuristics as well as metaheuristics. In the context of assortment optimization, typically greedy and construction heuristics are proposed which try to exploit knowledge about the problem structure. Other types of heuristics are applied less often. Jagabathula (2016) is one of the few examples proposing a local search heuristic.
  - **Approximation algorithms** are optimization methods that provide an approximate solution with guaranteed solution quality. That is to say it is possible to provide a bound on the quality of the returned solution for approximation algorithms. Note that if it is possible to formulate a bound on the solution quality, a heuristic turns into an approximation algorithm. We distinguish different types of approximation algorithms (see Schuurman and Woeginger 2009):

    1. An approximation algorithm is called *constant factor approximation (APX)* if it guarantees a constant approximation ratio and its running time is bounded by a polynomial in the problem size n. The corresponding complexity class APX includes all problems for which a polynomial time approximation algorithm with constant approximation ratio bound exists.
    2. Similarly, an algorithm is called *approximation scheme* for an optimization problem if it returns an output that is at least $1 - \epsilon$ times the optimal solution value and at most $1 + \epsilon$ times the optimal solution value, where $0 < \epsilon < 1$ denotes an arbitrary accuracy parameter.
    3. Such an approximation scheme is called *polynomial time approximation scheme (PTAS)*, if its computational complexity is polynomial in the instance size n for every fixed ϵ.
    4. Likewise, an approximation scheme is called *fully polynomial time approximation scheme (FPTAS)*, if its computational complexity is polynomial in n and $1/\epsilon$.

Note that both groups of solution concepts—exact and non-exact ones—each comprise an extremely broad variety of individual solution methods and combinations thereof. This also holds for the selection of solution methods that finds application in the area of assortment optimization. The solution approaches considered in the literature on

assortment optimization are evaluated in this review and stated in the literature classification provided in Tables 2, 3, 4, 5, and 6. Due to the sheer variety and broadness of the applied methods—which also can not be further classified in a meaningful way—for brevity we refrain from introducing each of them individually and refer the interested reader to the related articles for more information on the solution method of interest.

4.2 Performance guarantees and empirical evaluation

Research has recently focused on approximation-based approaches that provide performance guarantees. These guarantees can be denoted in various ways depending on the underlying method. In the following, we briefly introduce and exemplify different ways of providing performance guarantees.

- **Constant factor notation:** A popular way of denoting constant factor approximations is to simply provide a constant factor α implying that at least an approximation ratio of α can be obtained. An example for this notation is e.g. provided in Zhang et al. (2020) who obtain a 0.6 performance guarantee implying that their proposed approach is guaranteed to obtain at least three fifth of the optimal total expected revenue. Likewise, Udwani (2021) gives a 0.25 approximation algorithm. Alternatively, instead of providing the constant factor approximation ratio α, it is also possible to denote the performance guarantee as a percentage value $\alpha'\%$, meaning that the approximate solution is guaranteed to be at least $\alpha'\%$ of the optimal solution. This notation is e.g. used in Rusmevichientong et al. (2020) who provide a performance guarantee of 50% implying that their proposed approach is guaranteed to obtain at least 50 percent, i.e. half, of the optimal total expected revenue. However, in this context only few authors state whether their constant factor guarantee belongs to class APX.
- ϵ**-notation:** A performance guarantee is often denoted by $1-\epsilon$ implying that the output is at least $1-\epsilon$ times the optimal solution value, where $\epsilon > 0$ represents an arbitrary accuracy parameter. This notation is typically applied for approximation schemes such as PTAS or FPTAS. Examples for this notation are given in Feldman and Segev (2022) and Feldman and Topaloglu (2017b), who propose a PTAS, respectively, FPTAS with $1-\epsilon$ performance guarantee.
- **Big-oh notation:** Another way of denoting performance guarantees is the big-oh notation, which is applied for brevity and hides absolute constants. This notation is typically used in the literature on assortment optimization in case a dynamic problem is considered. The problem itself usually constitutes a regret minimization task, where regret refers to the gap between the expected revenue obtained by the proposed approach and the expected revenue according to an oracle with perfect information. An example for this notation is provided in Peeters and den Boer (2022), who obtain a performance guarantee of $O(\sqrt{T})$, implying that the performance of their proposed approach scales with the length of the selling period and is bounded by $\sqrt{T}$ times some constant factor.

Such theoretical performance guarantees—independent of the way they are denoted—only provide worst-case performance bounds. However, in practice the proposed solution approaches might perform way better than their theoretical worst-case guarantees.

This can be examined by evaluating a methods empirical performance, i.e. the performance of the proposed solution method compared to the optimal solution when being applied to synthetic or real data. We report the empirical performance of all articles considered in this review in Sect. 5.

Please note that the provided empirical performances are hardly comparable across different articles due to various reasons. First, different articles utilize different data sets—be it real or synthetic ones—with differing complexity and of different instance sizes. Moreover, different articles make use of different solvers and/or programming languages for implementing their proposed approaches. Finally, typically a limit on the maximum computation time for executing the proposed algorithm is set. However, these limits differ across the reviewed articles implying that there exist differences regarding whether and when an algorithm is enforced to stop. The earlier an algorithm is enforced to stop, the higher the probability that the best possible revenue that is obtainable by this algorithm is not found before stopping.

All these factors impact the empirical performance of the approaches proposed in the reviewed literature and thus limit their comparability. Thus, this indication of empirical performance is rather meant to gain an impression of whether the respective approach is exact or not but not meant to provide an exact performance that can be expected whenever the proposed approach is applied.

5 Classification of literature

This section is targeted to provide an extensive, structured overview of literature on pure assortment optimization. We expound the procedure of conducting our systematic literature review in Sect. 5.1. The description of the tabular presentation of our literature classification is provided in Sect. 5.2. Publications studying the assortment problem under parametric choice models are summarized in Sect. 5.3; studies on assortment optimization under nonparametric choice models are assembled in Sect. 5.4.

5.1 Systematic literature review procedure

As indicated by the title of this article, we conduct a systematic literature review. According to Thomé et al. (2016) and Durach et al. (2017), a systematic literature review—in contrast with a narrative one—follows a well-defined, replicable, scientific, and transparent process to identify, collect, appraise, and synthesize all relevant literature that meets certain predefined inclusion criteria to answer a specific research question and reports the evidence in a way that allows for clear conclusions regarding what is known and what is not known.

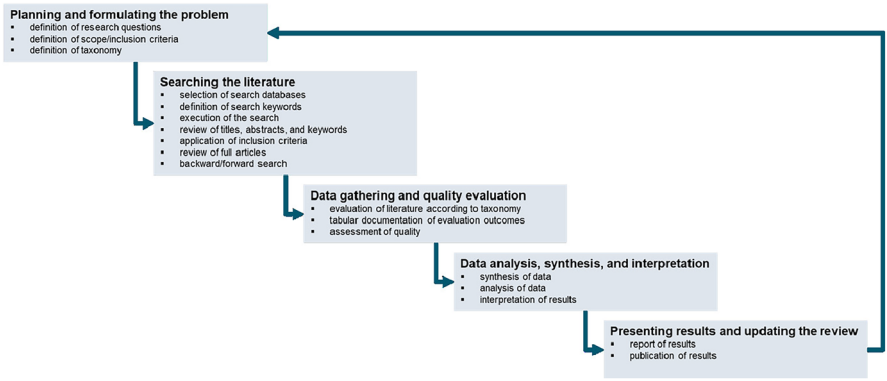

Fig. 3 Main steps and corresponding sub-tasks of our systematic literature review procedure; adapted from Thomé et al. (2016)

Thomé et al. (2016) provide a step-by-step approach for conducting such a systematic literature review in the context of operations management. Likewise, Durach et al. (2017) propose a step-by-step paradigm for systematic literature reviews in the context of supply chain management. Both guidelines basically comprise the same key components. For conducting our systematic literature review, we follow the five building blocks proposed by Thomé et al. (2016). To be precise, our review procedure consists of the five main steps and their corresponding sub-tasks visualized in Fig. 3. We briefly comment on each of the main steps in the following.

5.1.1 Planning and formulating the problem

In our review, we are interested in classifying the existing literature on assortment optimization according to a suitable taxonomy to determine research gaps in this research area. To identify the relevant literature, we select our inclusion criteria according to the following guideline. As mentioned before, our review is mainly targeted to the area of assortment optimization with a focus on pure assortment problems; related areas such as the extension to joint assortment and pricing problems are briefly introduced in Appendix B. Moreover, we restrict our review to mostly consider approaches that are based on choice modelling to capture consumer demand. To ensure that the studies contained in our review provide an application guideline, we limit our review to articles containing exemplary numerical studies. Finally, we restrict our review to the selection of articles that are published by the end of April 2023.

5.1.2 Searching the literature

To assemble an extensive set of literature on assortment optimization that satisfies our criteria, we select a variety of scientific databases and define a range of search keywords related to assortment optimization. To be precise, we choose Scopus, Science Direct, Springer, ACM Digital Library, IEEE Xplore, and Google Scholar as scientific databases and define two groups of search keywords. On the one hand,

we consider a selection of keywords related to the kind of optimization problem, namely 'assortment optimization', 'assortment personalization', and 'assortment planning'. On the other hand, we consider keywords covering the methodological component. These keywords comprise 'data-driven', 'parametric', 'nonparametric', and 'Machine Learning'.

The keywords are combined by using one keyword per group. The resulting combined keywords are used as search strings in the pre-selected scientific databases. As the databases ACM Digital Library and IEEE Xplore only yield very few hits, we change the search string to only using the search keywords of the first group, namely 'assortment optimization', 'assortment personalization', and 'assortment planning', for these databases. The resulting hits are pre-selected by title, abstract, and keywords if available. This yields a total of 309 articles. These publications are subsequently screened in further detail to ensure that they indeed fit the scope of our review. Moreover, we extend our literature base by relevant publications that are cited within these papers or required for providing further details on the topics covered by this review. In total, we consider 184 publications for this review.

5.1.3 Data gathering and quality evaluation

During the literature screening process, we collect a selection of key information on each paper that are used to group the publications according to their content. This key information includes factors related to the optimization problem itself, the customer behaviour, applied solution concepts as well as information related to the numerical experiments executed in the publication at hand. To be precise, the assembled literature on assortment optimization can be categorized according to the following factors:

1. Optimization problem (see Sect. 2)
    (a) non-robust vs. robust problem formulation
    (b) static vs. dynamic problem formulation
    (c) considered constraints
2. Customer behaviour (see Sect. 3)
    (a) choice model
    (b) single- vs. multi-purchase behaviour
    (c) consider-then-choose approach
    (d) sales channel
3. Solution concept (see Sect. 4)

(a) problem type
(b) computational complexity
(c) solution method
(d) exact vs. non-exact method
(e) performance guarantee

4. Numerical experiments

(a) data type (synthetic vs. real data)
(b) preference type (stated vs. revealed preference data)
(c) number of items
(d) price/revenue of considered items
(e) computation time
(f) empirical performance

We use the above taxonomy to classify the assembled literature on assortment optimization and tabularly document it in Tables 2, 3, 4, 5, and 6 of Sect. 5. To be precise, in line with Thomé et al. (2016), each article corresponds to one row in one of these tables and the proposed taxonomy is transferred into the tables' columns as detailed in Sect. 5.2. While collecting the key information according to the taxonomy for each of the considered articles, the fit between the reviews' goal and the design of the taxonomy is frequently evaluated and the taxonomy adjusted if required.

5.1.4 Data analysis, synthesis, and interpretation

We analyse the classification of the assembled literature that is documented in Tables 2, 3, 4, 5, and 6 of Sect. 5 in two ways. First, we study the values of each table column—i.e. of each factor in our taxonomy—individually across all considered articles. Second, we analyse combinations of values of different table columns across the considered articles. Doing so, we particularly focus on the existence and frequency of the individual values and value combinations, respectively. The results of this analysis are documented and interpreted in Sect. 5.

5.1.5 Presenting results and updating the review

The above proposed approach of analysing and synthesizing the literature on assortment optimization allows us to identify settings that have not yet been studied. To be precise, all settings for which no value or combination of values of certain classification factors exists according to our evaluation can be deemed research gaps. We summarize the identified research gaps in Sect. 6. Finally, our results are presented to the research community by documenting them in this review article.

Table 1 Description of the columns of Tables 2, 3, 4, 5, and 6

Column name	Description
Reference	Reference to considered publication
Robust	Robust or non-robust optimization: R=robust, NR=non-robust
Static	Type of optimization problem: s=static, d=dynamic
Channel	Retail channel: online, offline, omni=omni-channel, unspec.=unspecified
Choice model	Underlying choice model
CTC	Usage of consideration sets
#Purchase	#Purchasable products: single=single-purchase, multi=multi-purchase
Constraint	Constraint(s) used in the optimization problem
Type	Type of optimization problem: LP=linear, BFLP=binary fractional linear, NLP=nonlinear, DP=dynamic program
Comp. compl.	Computational complexity of the problem
Sol. method	Method used for approaching the assortment problem
Exact sol.	Exact solution: y=yes, n=no, b=both
Guarantee	Type of performance guarantee for non-exact solutions
Data	Data for num. exp.: syn=synthetic, real=real, SP=stated pref., RP=revealed pref
#Item	min. to max. number of products considered in the numerical experiments
Price	min. to max. product prices / revenues in the numerical experiments
Time[s]	Computation time (in s) of the proposed approach for the largest instance
Emp. perf.	Empirical average or worst case performance in the numerical experiments

5.2 Table structure

We transfer the taxonomy proposed in Sect. 5.1 into table columns in order to be able to present the results of the evaluation of the literature according to this taxonomy in a clear and comprehensible manner. To be precise, Tables 2, 3, 4, 5, and 6 capturing the literature classification—whereby each article corresponds to one row in one of these tables—all comprise the columns listed in Table 1.

Many of the properties listed in Table 1—such as robust, static, channel, choice model, CTC, #purchase, constraint, solution method, and exact sol.—are already introduced within the previous sections. Others—such as the columns reference, type, #item, and price—are self-explaining. We briefly comment on the remaining columns, i.e. comp. compl., guarantee, data, time, and emp. perf. in the following.

The column comp. compl. indicates the computational complexity, i.e. in our case the amount of time required for solving the considered assortment problem. Typically, it is distinguished whether a problem is in complexity class P or NP (see e.g. Whitley 2013; Homer and Selman 2011). In practice, the complexity class P (polynomial) can be thought of as all problems that are deemed tractable, which means that they can be solved in reasonable—i.e. polynomial—computation time (Homer and Selman 2011). The complexity class NP (non-deterministic polynomial) is the set of problems that are solvable in polynomial time on a non-deterministic Turing

Table 2 Overview of literature on static assortment optimization under the MNL choice model

	optimization problem											numerical experiments				
ref.[1]	robust	static	channel	CTC	#purchase	constraint	type	comp. compl.	sol. method[2]	exact sol.	guarantee	data	#item	price	time[s]	emp. perf.
1	NR	s	online	✓	single	cardinality	NLP	NP-comp.	linearization + rev.-ordered	n	PTAS	syn, real(RP)	20–1047	0–1	<1200	>98%
2	NR	s	offline	–	multi	cardinality	BFLP	strongly NP-hard	adj. rev.-ordered, LP relaxation	n	constant	syn	25–150	0–1	–	>99%
3	NR	s	offline	–	single	–	BFLP	–	enum. over candidate assortments	y	–	syn	8	–	–	100%
4	R	s	unspec.	–	single	cardinality	–	–	LP reformulation + gradient desc. w. line search	n	reg(n)	syn	20–60	0.5–0.8	–	–
5	NR	s	unspec.	✓	single	capacity	DP	NP-comp.	ADP	n	FPTAS	syn	36	1–50	<6	>96%
6	NR	s	online	–	single	cardinality, capacity	BFLP	P	enum. + bisection search	y	–	real(RP)	100–2000	–	<0.2	100%
7	NR	s	unspec.	–	multi	TU	NLP	NP-hard	ADP	n	PTAS	syn	15–60	0–1	<3600	>99%
8	NR	s	unspec.	✓	single	cardinality	NLP	NP-comp.	enum. over candidate assortments	n	PTAS	syn, real(SP)	10–25	1–200	<12	>99%
9	NR	s	online	–	single	TU	–	NP-hard	rev.-ordered, LP for auxiliary MNL	b	–	syn	4–15	1–10	–	–
10	R	s	unspec.	–	single	cardinality	NLP	–	enum. + bisection search	y	–	syn	20–100	1–50	<2.4	100%
11	NR	s	unspec.	–	single	cardinality, precedence	BFLP, LP	NP-hard	MILP, LP relaxation	n	constant	syn, real(RP)	100–200	–	<0.01	>99%
12	NR	s	unspec.	–	single	cardinality	NLP,LP	NP-hard	MILP	y	–	syn	100–1000	0–2000	<2.3	100%
13	R	s	unspec.	–	single	–	–	P	rev.-ordered	y	–	syn	20	0–6000	–	100%
14	NR	s	omni	–	single	cardinality	NLP, BFLP	P, NP-hard	bisection search + DP, geom. grid + DP	b	FPTAS	syn	16–64	1–10	<186	>95%
15	NR	s	online	–	single	–	NLP	NP-hard	rev.-ordered, construction heuristics	b	–	syn, real(RP)	10–100	0–1	<23	>92%
16	NR	s	unspec.	✓	single	cardinality	LP	–	MILP	y	–	real(SP)	117	–	–	100%
17	NR	s	online	✓	single	cardinality	NLP	–	construction heuristics	n	PTAS	real(RP)	20–900	0–30	<1.4	–
18	NR	s	unspec.	–	single	TU	NLP	–	LP, preference weight rounding	b	FPTAS	syn, real(RP)	60	0–10	<8	>99%
19	NR	s	online	–	single	cardinality	BFLP	–	binary search + MILPs	n	–	syn, real(RP)	100–50000	–	<16	>83%
20	NR	s	unspec.	–	single	–	NLP	NP-hard	rev.-ordered, quasi-rev.-ordered	b	–	real(RP)	4	0–5	–	100%
21	NR	s	unspec.	✓	single	capacity	NLP	NP-hard	rev.-ordered, k-quasi-attractiveness-ordered, DP	b	FPTAS	syn	5–20	1–20	–	>74%
22	NR	s	unspec.	–	single	–	NLP	NP-comp.	margin-ordered, NLP relaxation	b	–	real(RP)	–	–	<2.8	100%
23	NR	s	unspec.	–	multi	–	NLP,LP	–	MILP	y	–	syn, real(RP)	12–20	10–30	–	100%
24	NR	s	offline	✓	single	–	NLP	NP-comp.	quasi-markup-ordered, DP, ADP	b	FPTAS	syn, real(RP)	–	–	–	–

Table 2 (continued)

[1]references in this column: 1 = Aouad et al. (2021); 2 = Chen et al. (2022); 3 = Cachon et al. (2005); 4 = Dong et al. (2023); 5 = Feldman and Topaloglu (2017b); 6 = Feldman et al. (2021); 7 = Bai et al. (2023a); 8 = Bai et al. (2023b); 9 = Gallego and Berbeglia (2022); 10 = Hu et al. (2022); 11 = Kunnumkal and Martínez-de-Albéniz (2019); 12 = Leitner et al. (2023); 13 = Li and Ke (2019); 14 = Lo and Topaloglu (2021); 15 = Maragheh et al. (2021); 16 = Miller et al. (2010); 17 = Mushtaque and Pazour (2022); 18 = Sumida et al. (2020); 19 = Tulabandhula et al. (2022); 20 = Wang and Wang (2016); 21 = Wang and Sahin (2017); 22 = Wang (2021); 23 = Wang et al. (2023d); 24 = Wang et al. (2022b)

[2]Abbreviations used in this column: LP = Linear Programming; MILP = Mixed-Integer Linear Programming; NLP = Nonlinear Programming; DP = Dynamic Programming; ADP = Approximate Dynamic Programming; rev.-ordered = revenue-ordered; adj. rev.-ordered = adjustedrevenue-ordered; enum. = enumeration; gradient desc. w. line search = gradient descent with line search

machine (Whitley 2013). It is typically assumed that P≠ NP (Whitley 2013; Homer and Selman 2011). Following Whitley (2013), Homer and Selman (2011), and Schuurman and Woeginger (2009) a problem is said to be

1. NP-hard if it is at least as hard as any other problem in NP,
2. NP-complete if it is NP-hard and in NP,
3. strongly NP-hard if it remains NP-hard when all of its input parameters are bounded by a polynomial in the length of the input,
4. strongly NP-complete if it remains NP-complete when all of its input parameters are bounded by a polynomial in the length of the input,
5. APX-hard if there exists a PTAS reduction from every problem in APX to this problem,
6. APX-complete if the problem is APX-hard and in APX.

NP-hard problems cannot be solved in polynomial time. However, some NP-hard problems can be approximated in polynomial time—be it up to some constant approximation ratio (APX) or up to any approximation ratio (PTAS, FPTAS).

The column guarantee provides the type of performance guarantee—if any—that is given for non-exact solution approaches. To be precise, in this column we report whether a constant factor approximation, a PTAS or FPTAS, or a regret performance bound in big-oh notation is provided. That is to say, e.g. ϵ-approximations are only reported in this column in case they result from a PTAS or FPTAS. There are two publications providing an ϵ-approximation that does not result from a PTAS or FPTAS, namely Tulabandhula et al. (2022), whose guarantee is not polynomial at all and Chen and Jiang (2020b), who provide a pseudo-polynomial algorithm. In addition, we only report bounds in big-oh notation in case a regret optimization problem is considered. The notation reg$(\cdot)$ implies that a regret performance guarantee in big-oh notation is provided that depends at least on the parameters listed within the brackets but might depend on further parameters as well. Moreover, note that in case several types of guarantees are reported, the authors have either presented several solution methods or considered several different cases.

The column data captures two types of information regarding the data that are used for the numerical experiments. First, we distinguish whether synthetic or real data are explored. In the former case, the considered instances are typically

Table 3 Overview of literature on dynamic assortment optimization under the MNL choice model

	optimization problem											numerical experiments				
ref.[1]	robust	static	channel	CTC	#purchase	constraint	type	comp. compl.	sol. method[2]	exact sol.	guarantee	data	#item	price	time[s]	emp. perf.
1	NR	d	online	–	single	cardinality, partition,TU	NLP	–	extended UCB	n	$\text{reg}(n,T)$	syn, real	10–1728	1	–	–
2	NR	d	online	–	single	cardinality	NLP	–	decision rules from bandit algorithms	y	–	real(RP)	19	–	<32	100%
3	NR	d	online	–	single	cardinality	LP	–	MILP	n	–	real(RP)	19	–	150	>99%
4	NR	d	online	–	single	–	DP	NP-hard	myopic heuristic, one-step-lookahead heuristic	n	–	real(RP)	–	–	–	–
5	R	d	online	–	single	cardinality	–	–	active-elimination algorithm, fully-adaptive policy	n	$\text{reg}(n,T,C)$, $\text{reg}(n,T)$	syn	100–300	0.1–1	–	≥94%
6	NR	d	online	–	single	cardinality	NLP	–	Thompson Sampling	n	$\text{reg}(n,T,C)$	syn	20–150	0.1–1	–	–
7	NR	d	online	–	single	capacity	NLP	NP-hard	DP, enum. + approx. of reduction to known problem	n	PTAS	syn, real(RP)	50–100	1–5000	<1714	–
8	NR	d	unspec.	–	single	–	–	P	rev.-ordered	y	–	syn	5–50	0–10	<0.6	100%
9	NR	d	omni	–	single	cardinality, capacity	NLP, DP	P, NP-hard	rev.-ordered, DP, DP approx.	b	FPTAS	syn, real(RP)	10–20	0–1	<2220	>95%
10	NR	d	online	–	single	–	NLP	–	structure-aware algorithm	n	–	syn	100–400	–	–	–
11	NR	d	unspec.	–	single	cardinality, capacity	NLP	NP-comp.	DP + enum.	n	FPTAS	syn, real(SP)	18	0.3–1	<20	>96%
12	NR	d	unspec.	–	single	cardinality, inventory	–	–	primal-dual opt. + UCB + opt. solver	n	$\text{reg}(n,T)$, $\text{reg}(n,T,C)$	syn	10	0–1	–	–
13	NR	d	online	–	single	–	–	–	random projection + conv. opt. + UCB	n	–	syn, real(RP)	10–100	0–1	<376	–
14	NR	d	unspec.	–	single	–	NLP	–	stochastic approx.	n	$\text{reg}(T)$	syn	25–5000	0–1	–	–
15	NR	s,d	online	–	single	cardinality	NLP	–	enum. over candidate assortments + golden ratio search	b	–	real(RP)	200	–	–	–
16	R	s,d	unspec.	–	single	cardinality	NLP, DP	P	rev.-ordered	y	–	syn	10–20	175–600	–	100%
17	NR	d	unspec.	✓	single	inventory	DP	–	ADP	n	constant	syn	6	10–25	4.1	–
18	NR	d	unspec.	–	single	cardinality	DP	–	rev.-ordered	y	–	syn	3–10	175–600	–	100%
19	NR	s	unspec.	–	single	–	NLP	–	quasi-markup-ordered, LP	y	–	real(RP)	4	–	–	100%

Table 3 (continued)

[1]references in this column: 1 = Agrawal et al. (2019); 2 = Bernstein et al. (2019); 3 = Bernstein et al. (2022); 4 = Besbes et al. (2015); 5 = Chen et al. (2023a); 6 = Cheung and Simchi-Levi (2017); 7 = Feldman and Segev (2022); 8 = Flores et al. (2019); 9 = Gao et al. (2021); 10 = Kallus and Udell (2020); 11 = Liu et al. (2020); 12 = Miao et al. (2021); 13 = Miao and Chao (2022); 14 = Peeters and V. den Boer (2022); 15 = Rusmevichientong et al. (2010); 16 = Rusmevichientong and Topaloglu (2012); 17 = Rusmevichientong et al. (2020); 18 = Talluri and van Ryzin (2004); 19 = Wang (2018)

[2]Abbreviations used in this column: LP = Linear Programming; MILP = Mixed-Integer Linear Programming; DP = Dynamic Programming; ADP = Approximate Dynamic Programming; conv. opt. = convex optimization; UCB = Upper Confidence Bound (Bandit Algorithm); rev.-ordered = revenue-ordered; enum. = enumeration; approx. = approximation; opt. solver = optimization problem solver (standard solver)

generated by sampling from pre-defined distributions, whereas in the latter case, real-world data are gathered. Second, when real data are considered, we differentiate between stated and revealed preference data as introduced in Sect. 3.1. Hence, in Tables 2, 3, 4, 5, and 6 the differentiation between stated and revealed preferences is only provided when real data are considered. In this case, the information is stated in brackets.

The property time refers to the computation time measurement (in seconds) for executing the approach proposed in the respective paper for its largest considered instance. Please note that the provided computation times are hardly comparable across different articles due to several reasons. First, different articles utilize different resources e.g. in terms of hardware, processor, or memory. Moreover, different programming languages and solvers are applied for the execution of the numerical studies across different articles. Finally, different articles consider different instance sizes for their numerical studies. All these factors impact the execution time of the approaches proposed in the reviewed literature and thus limit their comparability. Thus, this indication of computation time is rather meant to gain an impression of whether the respective approach is very fast or very slow but not meant to provide an exact computation time that can be expected whenever the proposed approach is applied.

The column emp. perf. contains the empirical performance of the proposed solution approach as explained in Sect. 4.2. In case the column exact sol. indicates that only non-exact solution approaches are proposed, the authors typically determine the empirical performance by applying an existing method for finding the optimal solution or an upper bound thereof and comparing the solution obtained by their proposed approach with this exact solution or the upper bound thereof.

5.3 Parametric approaches

This section provides an overview of the literature on assortment optimization approaches whose underlying demand model belongs to the class of parametric choice models. The literature is separated by the underlying choice model. To be precise, Tables 2 and 3 contain the literature on static and dynamic assortment optimization with underlying multinomial logit choice model, respectively; Table 4

Table 4 Overview of literature on assortment optimization under MMNL, LC-MNL, NL, and PCL choice models

	optimization problem												numerical experiments				
ref.[1]	robust	static	channel	choice model	CTC	#purchase	constraint	type	comp. compl.	sol. method[2]	exact	sol. guarantee	data	#item	price	time[s]	emp. perf.
1	NR	s	unspec.	mix of indep. dem. & MNL	–	single	capacity	LP	NP-comp.	LP, DP + approx. auxiliary program, LP relaxation	b	FPTAS	real (SP)	10	–	–	–
2	NR	s	unspec.	LC-MNL	✓	single	–	–	NP-hard	rev.-ordered, adj. rev.-ordered	n	–	syn	5–100	1–100	–	–
3	NR	d	unspec.	LC-MNL	–	single	cardinality	DP	APX-hard, NP-hard	greedy, dependence-relax. approx., attraction value rounding + DP	n	FPTAS	syn	5–18	0–10	<0.0157, <0.0574	$\geq 99.4\%$, $\geq 99.9\%$
4	NR	s	unspec.	LC-MNL	✓	single	cardinality	NLP,LP	NP-hard	branch-and-cut	b	–	syn	500	100–350	<273	$>98\%$
5	NR	s	unspec.	MMNL	–	single	cardinality	BFLP	–	conic quadratic mixed 0–1	b	–	syn	200	1–3	<180	$>99.9\%$
6	NR	d	online	MMNL	–	single	inventory	BFLP	NP-comp.	threshold-policy, ADP	n	–	real (RP)	4–20	200–1000	–	$>97\%$
7	NR	s	unspec.	MMNL	–	single	capacity	BFLP	NP-comp.	geom. grid + approx. via continuous knapsack	n	–	syn	100	0–2000	<6	$>99\%$
8	NR	s	unspec.	MMNL	–	single	–	NLP	APX-hard	rev.-ordered	b	–	real (SP)	20–90	–	–	$>97\%$
9	NR	s	offline	MMNL	–	single	capacity	BFLP	NP-hard	mixed-integer second-order cone programming	y	–	syn	200–500	1–3	<189	100%
10	NR	s	unspec.	MMNL	–	single	–	NLP,DP	NP-comp.	Lagrangian relaxation, NLP relaxation, sub-gradient search	n	–	syn	50	0–200	–	–
11	NR	s,d	unspec.	MMNL	–	single	–	NLP,DP	NP-comp.	rev.-ordered	b	constant	syn	10–50	1–1000	–	$>96\%$
12	NR	s	unspec.	MMNL	✓	single	capacity	BFLP, LP	NP-hard	conic quadratic mixed 0-1 program + McCormick	n	–	syn	200–500	1–3	<175	$>75\%$
13	NR	s	unspec.	NL	–	single	–	NLP	NP-comp.	fractional prog. + branch-and-bound	y	–	syn	10–5000	0–3000	<70	100%
14	NR	s	unspec.	NL	–	single	capacity	NLP	NP-comp.	fixed-point problem via binary search	n	–	syn	7	49–99	–	–
15	NR	d	unspec.	NL	–	single	–	NLP	–	rev.-ordered + UCB	n	reg(T)	syn	500–10000	0.2–0.8	–	–
16	NR	s	unspec.	NL	–	single	–	LP	NP-hard	LP (+ candidate assort)	b	constant, PTAS	syn	125	1–10	–	$>94\%$
17	NR	s	unspec.	NL	–	single	cardinality, capacity	LP	P,NP-hard	LP (+ candidate assort)	b	constant	syn	45–150	0–12.5	<4	$>98\%$
18	NR	s	unspec.	NL	–	single	cardinality, capacity	LP	P,NP-hard	LP (+ candidate assort)	b	constant	syn	75–150	0–12.5	–	$>95\%$
19	NR	s,d	unspec.	NL	–	single	–	NLP	P	construction heuristic	y	–	syn	9	0–15	–	100%
20	NR	d	unspec.	NL	–	single	–	DP	–	DP	y	–	real (RP)	8	1.3–7.8	–	100%
21	NR	s	unspec.	NL	–	single	cardinality	NLP	P	enum., DP, binary search	y	–	syn	10–500	0–12.5	–	100%
22	NR	s	unspec.	PCL	–	single	capacity	BFLP	NP-hard	fixed-point problem via LP relaxation	n	constant	syn	30–60	0–1	–	$>94\%$
23	NR	s	unspec.	PCL	–	single	cardinality, capacity, partition	BFLP	NP-hard	binary search, LP/pipage rounding, SDP/LP relaxation	n	constant, PTAS	syn	30–100	0–1	<90	$>96\%$
24	NR	s	unspec.	PCL	–	single	cardinality	BFLP	strongly NP-hard	fixed-point problem via LP/SDP relaxation + random/iterative rounding	n	constant	syn	50–100	0–1	<0.25	$>95\%$

Table 4 (continued)

[1]references in this column: 1 = Cao et al. (2022); 2 = Berbeglia et al. (2021a); 3 = Gallego et al. (2023); 4 = Méndez-Díaz et al. (2014); 5 = Atamtürk and Gómez (2020); 6 = Bernstein et al. (2015); 7 = Feldman and Topaloglu (2015a); 8 = Goutam et al. (2020); 9 = Jiang and Nip (2022); 10 = Kunnumkal (2015); 11 = Rusmevichientong et al. (2014); 12 = Şen et al. (2018); 13 = Alfandari et al. (2021); 14 = Chen and Jiang (2020b); 15 = Chen et al. (2021b); 16 = Davis et al. (2014); 17 = Feldman and Topaloglu (2015a); 18 = Gallego and Topaloglu (2014); 19 = Li et al. (2015); 20 = Qiu et al. (2020); 21 = Xie and Ge (2018); 22 = Feldman (2017); 23 = Ghuge et al. (2021); 24 = Zhang et al. (2020)

[2]Abbreviations used in this column: LP = Linear Programming; NLP = Nonlinear Programming; fractional prog. = Fractional Programming; SDP = Semidefinite Programming; DP = Dynamic Programming; ADP = Approximate Dynamic Programming; rev.-ordered = revenue-ordered; adj. rev.-ordered = adjusted revenue-ordered; approx. = approximate; dependence-relax. approx. = dependence relaxation approximation; geom. grid = geometric grid; assort = assortments; enum. = enumeration

captures the literature on assortment optimization with underlying logit choice models such as MMNL, LC-MNL, NL, and PCL, and Table 5 assembles the literature on assortment optimization under further parametric choice models.

The literature captured by each table is classified according to a selection of key factors related to the optimization problem itself and to the numerical experiments executed in the publication at hand. The former factors comprise information whether a robust approach is considered, an indication whether the problem is static or dynamic, the sales channel, the underlying choice model, whether a consider-then-choose approach is applied, whether a single- or a multi-purchase setting is considered, incorporated constraints, the problem type, the computational complexity, the solution approach, its exactness, and provided performance guarantees. The key factors related to the numerical experiments cover an indication whether synthetic or real (stated or revealed preference) data are analysed, the number of products used in the numerical experiments, the price or revenue range of the considered products, the computation time for the largest considered instance as well as the empirical average or worst case performance of the proposed approach. In summary, the columns of the subsequent tables cover the content listed in Table 1. Based on the selection of an appropriate assortment optimization setting using the criteria from Sect. 2, a suitable choice model from Sect. 3, and an appropriate solution concept as addressed in Sect. 4, researchers and practitioners can easily identify matching studies from Tables 2, 3, 4, and 5 according to their properties.

Overall, we consider 82 studies on assortment optimization under a parametric choice model. Across all these publications, the researchers typically seem to first consider the pure static assortment problem under a certain choice model, subsequently extend it by cardinality and capacity constraints followed by more complex constraints, before considering dynamic or robust versions of the assortment problem. Due to its simplicity, particularly the MNL model is a popular choice for introducing new settings. Hence, there exists by far more literature on assortment optimization under the MNL model compared to other choice models. To be precise, among the 82 publications on assortment optimization under parametric choice models, 43 studies utilize the MNL as underlying choice model, followed by 9

Table 5 Overview of literature on assortment optimization under further parametric choice models including exponomial choice, Markov chain choice, single transition, locational choice, probabilistic choice, attraction demand, and general choice models

	optimization problem												numerical experiments				
ref.[1]	robust	static	channel	choice model	CTC	#purchase	constraint	type	comp. compl.	sol. method[6]	exact	sol. guarantee	data	#item	price	time[s]	emp. perf.
1	NR	s	unspec.	EXP	–	single	cardinality, capacity,TU	DP	NP-hard	ADP	n	FPTAS	syn, real (SP,RP)	10	–	–	>97%
2	NR	s	unspec.	EXP	–	single	–	–	–	construction heuristic	n	–	syn, real(RP)	10–15	–	–	>97%
3	NR	s	unspec.	MCC	–	single	–	LP	P	DP, LP	y	–	syn	10–100	1.1–1.1^{100}	–	100%
4	R	s	unspec.	MCC	–	single	–	LP	P, APX-hard	min-max duality, fixed-point problem	y	–	syn	10–50	0–1	<20	100%
5	NR	s	online	MCC	✓	single	cardinality, capacity	–	APX-hard	incremental greedy, iterative externality-adjust, linearized exp. revenue function	n	constant	syn	30–200	0–1	<48, <464	>77%
6	NR	s,d	unspec.	MCC	–	single	cardinality	LP	P	LP, LP approx.	b	–	syn	100	0–100	<1	100%
7	NR	s,d	unspec.	MCC	–	single	–	LP	P	construction heuristics	b	reg(n, T)	syn	5–100	–	–	>98%
8	NR	s	online	single transition[2]	–	single	–	NLP	NP-hard	rev.-ordered, DP, max. CSP + SDP rounding, submodularity, MILP	b	constant	syn	5–500	0–1	<32, <0.1	100%, >67%
9	NR	s	unspec.	locational choice[3]	–	single	–	–	–	genetic algorithm, simulated annealing, tabu search	n	–	syn	–	–	<1	>96%
10	NR	s	unspec.	locational choice	–	single	–	–	–	DP + line search	y	–	syn	–	–	<1	100%
11	NR	d	unspec.	probab. choice[4]	–	single	capacity	–	–	stochastic approx., UCB, bisection method	n	–	syn	29–64	–	–	–
12	NR	d	unspec.	attraction demand[5]	–	single	cardinality, capacity	BFLP	NP-comp.	NLP relaxation, construction heuristics	n	constant	syn, real(RP)	4–150	–	–	>90%
13	NR	d	online	general choice	–	single	capacity	–	–	LP-based heuristic	n	constant	syn, real(RP)	73	9–81	–	>62%
14	NR	d	online	general choice	✓	single	–	–	–	myopic policy	n	constant	syn	5	15–30	–	>85%
15	NR	s	unspec.	general choice	✓	single	cardinality	–	–	local search heuristic	b	–	syn	100–200	0–2000	–	>99%

Table 5 (continued)

[1]references in this column: 1 = Aouad et al. (2022); 2 = Alptekinoğlu and Semple (2016); 3 = Blanchet et al. (2016); 4 = Désir et al. (2023); 5 = Désir et al. (2019); 6 = Feldman and Topaloglu (2017a); 7 = Gallego and Lu (2021); 8 = Nip et al. (2021); 9 = McElreath et al. (2010); 10 = McElreath and Mayorga (2012); 11 = Peeters et al. (2022); 12 = Caro et al. (2014); 13 = Golrezaei et al. (2014); 14 = Gong et al. (2021); 15 = Jagabathula (2016)

[2]The single transition model is similar to the MCC model. However, under the single transition choice model the seller can recommend a subset of available products if the customer arrives at a non-available one. Moreover, this model assumes that a customer either purchases a product or leaves after a single transition

[3]The locational choice model is a utility-based model where products are viewed as a bundle of their attributes. The set of product attributes is specified by the products' location. The firm can control the rate of substitution between products by choosing their locations relative to each other

[4]The probabilistic choice model forms the continuous counterpart of the widely studied discrete multinomial logit model

[5]In theattraction demand model, each product's market share contribution is assumed to be proportional to its preference weight or attractiveness in each period

[6]abbreviations used in this column: LP = Linear Programming; MILP = Mixed-Integer Linear Programming; NLP = Nonlinear Programming; SDP = Semidefinite Programming; DP = Dynamic Programming; ADP = Approximate Dynamic Programming; max. CSP = maximum Constraint Satisfaction Problem; UCB = Upper Confidence Bound (Bandit Algorithm); rev.-ordered = revenue-ordered; approx. = approximation; adjust = adjustment; exp. = expected

publications under the NL model, 8 studies under the MMNL model, and 5 publications under the Markov chain choice model.

Moreover, as can be seen in Tables 2, 3, 4, and 5, most authors analyse the non-robust assortment problem—only 6 out of 82 publications who study the assortment problem under a parametric choice model follow a robust approach. Out of these six studies, five consider the assortment problem under the MNL model; only one publication on robust assortment optimization utilizes the Markov chain choice model. Moreover, all of these studies on robust assortment optimization under a parametric choice model consider single-purchase settings without constraint or with cardinality constraint.

Similarly, the majority of the 82 studies deal with the static assortment problem; only 30 out of 82 studies consider the dynamic problem, though the fraction of researchers studying the dynamic version increases in recent years. This finding might be attributable to the increasing interest in online settings, where typically sequential results pages are considered. The publications considering dynamic assortment problems mostly assume that the customer demand follows a MNL model; only twelve studies analyse this setting for different parametric choice models, namely MMNL, LC-MNL, NL, MCC, probabilistic choice, attraction demand, and general choice model.

In addition, according to our literature overview, most of the 82 studies are not explicitly targeted to a certain sales channel. Among those studies that can be attributed to a certain channel, 21 study the assortment problem in an online setting, four explicitly deal with an offline setting, and only two publications consider an

Table 6 Overview of literature on assortment optimization under nonparametric choice models

	optimization problem												numerical experiments				
ref.[1]	robust	static	channel	choice model	CTC	#purchase	constraint	type	comp. compl.	sol. method[2]	exact sol. guarantee		data	#item	price	time[s]	emp. perf.
1	NR	d	unspec.	general rank-based	–	single	capacity	–	NP-hard	approx. static problem (+ price thresholding) + greedily opt. multi-item newsvendor problem	n	PTAS	syn	20	0–1	<12	–
2	NR	s	unspec.	consider-then-choose	✓	single	cardinality	DP	P	bipartite graph representation, DP, divide-and-conquer	y	–	syn, real(RP)	17–200	–	<139	100%
3	NR	s	unspec.	generic non-param.	–	single	precedence, partition	LP	–	MILP	y	–	syn	30	1–100	<7	100%
4	NR	s	unspec.	decision forest	–	single	–	LP	NP-hard	MILP, linear relaxation, Benders decomposition	n	–	syn	100–3000	1–100	<3601	>65%, >82%
5	NR	s	unspec.	mixture of Mallows	–	single	cardinality, capacity	LP	NP-hard	MILP, enum. of candidate assortments	n	PTAS	syn, real(SP)	10–100	0–1	≤166, <19	>92%
6	NR	s	unspec.	k-product non-param.	✓	single	cardinality	LP	strongly NP-hard	LP relaxation + rounding, SDP relaxation + rounding	n	constant	syn	50–100	–	<600	>96%
7	NR	s	unspec.	sequential flips	–	single	capacity	–	NP-hard	LP relaxation + DP	n	PTAS	syn	100	0–1000	<33	>99%
8	NR	s	unspec.	rank-based	✓	single	–	–	NP-hard	DP	y	–	syn	5–50	–	<9	100%
9	NR	s	unspec.	rank-based	✓	single	–	–	NP-hard	construction heuristic + candidate assortments	y	–	syn	4–50	1–20	<558	100%
10	NR	s	unspec.	partially ranked	✓	single	–	LP	NP-hard	MILP	y	–	syn, real(RP)	30–1000	–	<43260, <1020	100%
11	R	d	unspec.	general rank-based	–	single	cardinality, capacity	NLP	NP-hard	cutting-plane, greedy	b	–	syn	20–100	0–12.5	<1436	>95%
12	NR	s	unspec.	tree choice	✓	single	cardinality	DP	APX-hard	DP	y	–	syn, real(RP)	7–511	0–511	–	100%
13	NR	s	online	consider-then-choose	✓	single	–	LP	–	LP	y	–	syn	5–84	0–3	<260	100%

Table 6 (continued)

[1]references in this column: 1 = Aouad et al. (2018a); 2 = Aouad et al. (2020); 3 = Bertsimas and Mišić (2015); 4 = Chen and Mišić (2021); 5 = Désir et al. (2021); 6 = Feldman et al. (2019); 7 = Feldman and Paul (2019); 8 = Honhon et al. (2012); 9 = Honhon et al. (2020); 10 = Jena et al. (2020); 11 = Mehrani and Sefair (2022); 12 = Paul et al. (2018); 13 = Schwamberger et al. (2023)

[2]LP = Linear Programming; MILP = Mixed-Integer Linear Programming; SDP = Semidefinite Programming; DP = Dynamic Programming; approx. = approximate; opt. = optimize; enum. = enumeration

omni-channel setting. Most of these online settings as well as the omni-channel settings assume that consumer demand follows a MNL model.

Moreover, many online settings incorporate constraints—mostly cardinality and capacity constraints but also TU, partition, and inventory constraints. Finally, note that all of the publications specialized on online channels study the single-purchase version of the assortment problem. This finding does not only hold for the online setting. Overall, most authors consider single-purchase settings, though we observe an increasing interest in multi-purchase settings. The only three studies considering a multi-purchase setting assume that consumer demand follows a MNL model.

In addition, according to our literature overview, the vast majority of the studies either do not incorporate any constraint or focus solely on cardinality or capacity constraints though a variety of other constraints such as TU, precedence, partition, and inventory constraints are also considered.

Furthermore, 31 out of the 82 studies on assortment optimization under a parametric choice model do not report the computational complexity of their considered problem. Out of the 51 articles who do comment on the computational complexity, 14 state their problem to be in P, 23 claim their problem to be NP-hard, 14 articles mention that they study NP-complete problems, 2 studies indicate their problems to be strongly NP-hard, and 4 articles remark that they consider APX-hard problems. Note that some articles consider more than one problem at once (e.g. unconstrained and constrained settings) and thus might face different complexity classes.

Moreover, due to the high complexity of most of the assortment problem formulations, the majority of the authors do not provide exact solutions to the assortment problem, whereby 38 out of the 62 studies proposing non-exact approaches provide theoretical performance guarantees. Finally, we note that most of the studies, namely 68 out of 82 publications, do not take a consideration set into account. Only 14 studies analyse the assortment problem while following a consider-then-choose approach. All of these 14 studies consider non-robust, single-purchase settings. Moreover, most of them assume that consumer demand follows the MNL model though there also exist some studies considering this setting under LC-MNL, MMNL, Markov chain, or general choice models.

All of the 82 publications with underlying parametric choice model considered by us execute a numerical study to verify the practical applicability of their proposed approaches. Most of these studies, namely 68 out of 82, utilize synthetic data for their numerical experiments; 31 (additionally) consider real data. However, the data generation processes for creating synthetic data and the analysed real data sets vary heavily across the considered literature. Out of the 31 studies analysing real data, 25 consider revealed preferences, 5 have a look at stated preferences and one study utilizes

both, stated and revealed preference data. Furthermore, according to our literature overview, the lowest number of possible items to be included in the offer set of the numerical experiments is 3; the largest number of products considered is 50000. In addition, many studies assume a price or revenue range between 0 and 1. For the remaining studies, the considered prices or revenues range between 1 and 13780.

Moreover, according to our literature overview only 38 out of the 82 studies with underlying parametric choice model provide information regarding the computation time of their proposed approach, whereas the vast majority of the studies, namely 60 out of 82, document their methods' empirical performance. Finally, we notice that the empirically observed performance typically by far outperforms the theoretical guarantees.

5.4 Nonparametric approaches

This section provides an overview of the literature on assortment optimization whose underlying demand model belongs to the class of nonparametric choice models. The literature is summarized in Table 6 and classified according to a selection of key factors related to the optimization problem itself and to the numerical experiments executed in the publication at hand. The column descriptions for Table 6 are provided in Table 1.

Overall, we consider 13 studies on assortment optimization under a nonparametric choice model. Across all these publications, general rank-based choice models appear to be the most popular ones. To be precise, 5 out of 13 publications assume a rank-based model. Moreover, as can be seen in Table 6, the vast majority of the publications study the non-robust assortment problem; only one study analyses the robust version of the assortment problem under a nonparametric choice model. This publication studies a dynamic setting, is not targeted towards a specific sales channel, and does not incorporate a consideration set but allows for cardinality and capacity constraints.

Furthermore, according to our literature overview, most studies consider the static setting though there are also two publications analysing the dynamic version of the assortment problem. Likewise, we find that all studies on assortment optimization under a nonparametric choice model focus on the single-purchase setting and all except for one study are not targeted towards a specific sales channel. The publication targeted to a certain sales channel studies an online setting. Moreover, note that the studies considered by us incorporate a broad range of constraints including cardinality, capacity, precedence, and partition constraints.

Besides, we notice that only two out of the 13 studies on assortment optimization under a nonparametric choice model do not comment on the computational complexity of their considered problem. Out of the 11 articles who do comment on the computational complexity, one states the problem to be in P, whereas 8 studies consider NP-hard problems, one article mentions that an NP-hard problem is considered and one article reports that they study an APX-hard problem.

In addition, we find that according to Table 6, seven of the considered studies follow an exact solution approach, five studies apply non-exact solution methods,

and one study considers both—exact and non-exact solution approaches. Among those six studies considering a non-exact solution method, four provide a theoretical performance guarantee for the proposed solution procedure. Furthermore, we note that seven of the publications on assortment optimization under a nonparametric choice model take consideration sets into account.

All 13 publications with underlying nonparametric choice model considered by us execute a numerical study to verify the practical applicability of their proposed approaches. Interestingly, all studies base their numerical experiments on synthetic data; four of them additionally consider real data sets. Out of those 4 studies considering real data sets, 3 focus on revealed preference data, whereas only one study utilizes stated preference data. Furthermore, according to Table 6, the lowest number of possible items to be included in the offer set of the numerical experiments is 4, the largest number of products considered is 3000. In addition, note that the considered prices or revenues range between 0 and 1000.

Moreover, according to our literature overview all but one of the 13 studies with underlying nonparametric choice model provide information regarding the computation time and the empirical performance of their proposed approach. Finally, we notice that the empirically observed performance typically by far outperforms the theoretical guarantees.

6 Future research

This section is targeted to provide a structured overview of potential future research areas. These potential research areas comprise assortment problem settings that have not yet been studied according to our proposed taxonomy, assortment optimization under further demand models as well as intrinsic assortment optimization.

6.1 Unstudied settings according to the taxonomy

As explained in Sect. 5.1, our systematic literature review is targeted to identify research gaps within the research area of assortment optimization according to our proposed taxonomy. Based on the literature overview provided in Sect. 5, we observe a variety of assortment optimization settings under parametric and nonparametric choice models that are not yet studied. We start by proposing future research areas under parametric choice models followed by potential research areas under nonparametric ones.

As mentioned before, most assortment problems under parametric choice models are studied in non-robust settings. The publications studying the robust assortment problem typically assume that consumer demand follows a MNL model; only one study considers the robust assortment problem under a Markov chain choice model. Moreover, all these studies analysing the robust assortment problem consider single-purchase settings and do not take consideration sets into account. Hence, it

would be worth investigating the robust multi-purchase assortment problem and the robust assortment problem with consideration sets. In addition, the studies on robust assortment optimization under a parametric choice model either do not consider any constraint or incorporate a cardinality constraint. It might be worth studying this problem while accounting for further constraints. Finally, we notice that most studies on assortment optimization under parametric choice models do not consider an omni-channel setting. Publications, particularly studying online or omni-channel settings, typically assume that consumer demand follows a MNL model. It might be worth investigating this setting under further parametric choice models.

Besides these assortment problem settings under parametric choice models, we additionally identify further research areas under nonparametric choice models. For example, we find that only one publication on assortment optimization under a nonparametric choice model studies the robust assortment problem. To be precise, this publication considers a dynamic version of the problem. Hence, it would be worth to specifically investigate the static robust assortment problem under a nonparametric choice model. Moreover, only of the studies on assortment optimization under a nonparametric choice model is specifically targeted towards a certain sales channel—in this case the online sales channel. It would be worth investigating this setting for omni-channel environments as well. Finally, we notice that none of the studies analyses the assortment problem for multi-purchase settings. Hence, it would be worth studying the multi-purchase assortment problem under nonparametric choice models.

6.2 Assortment optimization under other demand models

Besides conducting research on assortment optimization under the previously specified settings, another area of future research is based on determining further choice models and demand modelling techniques and on addressing the resulting assortment optimization problems.

Over the past decades, an extreme boost in the application of machine learning techniques can be observed. This has also affected the literature on demand modelling. In recent years, researchers executed plenty of studies to compare the predictive performance of the choice models introduced so far with a variety of the most common machine learning models (see e.g. Wang et al. 2021 an the references therein). Most of the studies comparing the predictive ability of ML and choice models find that ML approaches by far outperform the classic choice models in terms of predictive accuracy. However, this does not imply that ML-based approaches are also superior for assortment optimization as it is difficult to optimize over ML models—particularly more sophisticated ones such as tree-based ensembles.

One method to use machine learning models for assortment optimization is to simply multiply the demand probabilities predicted by the ML model with the product revenues and take those products with highest probability times revenue values. Feldman et al. (2021) compare a classic MNL-based assortment approach with this ML-based one and find that the MNL-based method by far outperforms the machine

learning approach in terms of maximal revenue. However, note that this finding might be attributable to the fact that the applied ML-based approach strongly oversimplifies the problem as in this case the purchase probabilities do not depend on the set of offered products.

Likewise, Peng et al. (2022) investigate whether and how better prediction accuracy transforms into better decisions for assortment planning by comparing MNL, DeepFM, and a version of DeepFM that accounts for assortment information. The authors find that a choice model with better predictive power may not yield higher revenues. Hence, more work needs to be done in combining machine learning and assortment optimization to better exploit the superior predictive accuracy of machine learning-based demand models for assortment planning.

One research direction that has developed rapidly in recent years is the modelling of demand using deep learning-based approaches. Among them, Aouad and Désir (2022) propose a neural network-based choice model called RUMnet. This choice model is consistent with the RUM framework and formulates the random utility function using a sample average approximation method. The authors show that RUMnets are able to approximate any RUM choice model with arbitrary accuracy and find that their proposed model outperforms other state-of-the-art choice modelling and machine learning methods by a significant margin on two considered real-world data sets. It would be worth to investigate the assortment optimization problem under this choice model.

Cai et al. (2022) also develop deep learning-based choice models. To be precise, the authors study two settings of choice modelling—feature-free and feature-based—and propose neural network models that are able to capture both, the intrinsic utility for each candidate choice and the effect that the assortment has on the choice probability. The authors provide evidence that the proposed deep learning-based choice models are capable of recovering existing choice models with an effective learning procedure. Moreover, they find that such models are particularly useful in case the underlying model/training data are too complex to be described by a simpler choice model such as MNL and when there are sufficient training data (e.g. 5000 samples for 20–50 products). Following this, Wang et al. (2023a) propose a mixed-integer programming formulation for the corresponding assortment problem that is solvable by off-the-shelf integer programming solvers. However, since the approach is MIP based, the size of the optimization cannot scale to more than around a hundred products. Hence, further research and new optimization algorithms are required to scale beyond this.

Wang et al. (2023b) propose a transformer neural network architecture, the Transformer Choice Net, which does not only take customer and product features into account but also considers contextual information such as the offered assortment or the customer's past choices. By being able to predict multiple choices, this model is particularly suitable in situations where the customer chooses more than one item—such as in e-commerce shopping. The authors provide empirical evidence that their architecture beats leading models in the literature in terms of out-of-sample prediction performance on a range of benchmark data sets. Given the strong predictive power of Transformer Choice Nets, an interesting avenue to explore would be their application in assortment optimization.

Besides modelling demand using deep learning-based approaches, further choice models are recently developed. For example, Alptekinoğlu and Semple (2021) introduce the heteroscedastic exponomial choice model that generalizes the classic exponomial choice model by allowing the variance of the exponentially distributed random component of the utility to be product-specific. According to the study of Berbeglia et al. (2021a), the classic exponomial choice model exhibits an outstanding performance both in terms of predictive ability and in terms of revenue performance. Hence, it would be worth to investigate the assortment problem under the newly proposed heteroscedastic exponomial choice model.

Finally, there exists vast literature on dynamic discrete choice modelling, which is a natural extension of the static discrete choice modelling framework (see Keane and Wolpin 2009). Such dynamic discrete choice models are particularly designed for dealing with dynamic settings. A review on dynamic discrete choice models is e.g. provided by Aguirregabiria and Mira (2010). It might be interesting to study the dynamic assortment problem (Dynamic AOP) under such dynamic discrete choice models.

6.3 Intrinsic assortment optimization

Finally, we would like to draw attention towards an approach on optimization using machine learning techniques that is gaining increasing interest in recent times. The approach is based on the idea of intrinsic optimization. In this case, the feature to be optimized—e.g. the assortment or the price—is not assumed to be fixed in advance but is optimized while training the ML model itself. Mišić (2020) studies the tree ensemble optimization problem by answering the question "given a tree ensemble that predicts some dependent variable using controllable independent variables, how should we set these variables so as to maximize the predicted value?". The author formulates the problem as a mixed-integer optimization and shows that their methodology can efficiently solve large-scale problem instances to near or full optimality.

Since this approach scales exponential in runtime, Perakis and Thayaparan (2023) propose UMOTEM, an algorithm form solving a constrained optimization problem where the objective function is determined by a tree ensemble model. The proposed algorithm significantly reduces the problems' complexity since the number of binary variables only scales linearly instead of following an exponential growth. The authors demonstrate that their algorithm is able to capture more than 90% of optimality on a variety of data sets. One potential area of further research is to adapt this intrinsic optimization approaches to assortment problems.

7 Conclusion

Research on assortment optimization received a considerable boost in attention over the past decade. Various assortment problem settings under diverse choice models have been investigated with the aim of establishing efficient solution approaches.

However, due to the sheer amount of different approaches to assortment optimization available in operations research literature, it is difficult to keep track of all available ones. Our review supports the reader by providing an extensive overview of different available assortment optimization settings.

To be precise, we introduce different assortment optimization settings such as robust, non-robust, static, or dynamic assortment problems for different sales channels that might follow a consider-then choose approach, consider single- or multi-purchase settings and selected business constraints under a variety of different choice models and are solved using different solution concepts.

Based on this, we assemble an extensive literature overview on pure assortment problems under parametric and nonparametric choice models. The literature is classified according to a proposed taxonomy. Our taxonomy takes a selection of key factors related to the assortment problem itself, the customer choice behaviour, the solution concept as well as information related to the executed numerical experiments into account. This makes it easy for academics and practitioners alike to determine the assortment optimization setting that is most suitable for them and identify relevant related literature.

Finally, we conclude our review by outlining potential future research areas that deserve some attention but have barely been addressed in the literature so far. These potential research areas comprise a variety of assortment optimization settings that are not yet studied according to our literature overview but also include research areas related to determining new demand modelling approaches and solving the corresponding assortment problems as well as research on the topic of intrinsic optimization. We hope that this review spurs further research on assortment settings barely addressed so far and further propagates the research on and application of assortment optimization.

8 Supplementary information

The authors do not provide any supplementary materials.

Appendix A: Choice model design & estimation

In recent years, the assortment problem has been studied under a variety of choice models that are briefly introduced in Sect. 3. This section is targeted to provide a more detailed, formal introduction of the previously mentioned choice models. To be precise, in Section A.1 and Section A.2, we provide an overview of the most popular parametric, respectively, nonparametric choice models considered in the literature whereby we follow Strauss et al. (2018) in assigning the choice models to parametric and nonparametric approaches.

A.1 Parametric choice models

Parametric choice models are based on random utility theory, where it is assumed that consumers associate a certain utility with every product, and decide on the alternative that maximizes their utility (Strauss et al. 2018). This framework is referred to as random utility maximization (RUM). Within this framework, the utility $U_i = u_i + \epsilon_i$ of product i is composed of the deterministic part u_i and a random component ϵ_i. Using this, the probability $p_i(S)$ that product i is chosen among the offered assortment S is given by the probability that this product is associated with the highest utility, i.e.

$$p_i(S) = P\big(U_i \geq U_j \quad \forall j \in S \cup \{0\}\big).$$

The deterministic component u_i can be expressed as a linear function $u_i = \beta^T z_i$ of an attribute vector z_i that influences the purchase probabilities (see Strauss et al. 2018). Regarding the attributes influencing the deterministic part of the utility, one can distinguish between different types of determinants—individual-specific and alternative-specific ones. Individual-specific variables describe the characteristics of the decision maker such as income or age, whereas alternative-specific variables vary over both, individuals and alternatives. An example for the latter type of variable is the time an individual would need to travel with a certain travel mode. For more information on individual-specific and alternative-specific variables, we refer the interested reader to Heiss (2002).

Different parametric choice models result from different assumptions made on the distribution of the random component (Strauss et al. 2018).

A.1.1 Multinomial logit

The most popular parametric choice model is the multinomial logit (MNL) choice model of Luce (2012) and McFadden (1973). This model is particularly known for its simplicity and can be identified as member of the RUM framework by choosing the random components ϵ_i to be iid. random variables that follow the Gumbel distribution with a common scale parameter, typically normalized to one, and location parameters $u_i, i \in N$ with $u_0 := 0$. Under the MNL model, the probability to select a product i from the offer set S is determined by its utility relative to the total utility of the offer set; more formally:

$$p_i(S) = \frac{e^{u_i}}{1 + \sum_{j \in S} e^{u_j}}.$$

McFadden (1978) show that the parameters of this model can be estimated easily and Talluri and van Ryzin (2004) note that the corresponding optimization problem can be solved efficiently. To be precise, Talluri and van Ryzin (2004) prove that the optimal assortment under the MNL model is revenue-ordered, i.e. the optimal assortment consists of a number of products whose revenues are higher than the

revenues of those products that are not selected. As shown by Davis et al. (2013), this problem is even solvable under TU constraints.

Nevertheless, it should be taken into account that the MNL model might have a deficiency in representing the choice among alternatives with shared attributes—the Independence of Irrelevant Alternatives (IIA) property (see Ben-Akiva and Lerman 1985) illustrated by the well-known 'red bus/blue bus' paradox (Debreu 1960)—and should therefore be used with caution according to Talluri and van Ryzin (2004). Under the IIA property, substitution across alternatives is proportional, which can lead to the overestimation of choice probabilities for products that are considered similar by the customer, see Strauss et al. (2018). Hence, in case the choice set contains subgroups whose products are perceived more similar than products across different subgroups, the IIA property does not hold (Kök et al. 2008). Then, the MNL model might not be a suitable choice for modelling consumer demand.

A.1.2 Mixed multinomial logit

The mixed multinomial logit (MMNL) choice model (McFadden and Train 2000) considers different customer segments whereby the preferences of each segment $l \in L$ follow a segment-specific MNL model. We distinguish two cases. When the segment each customer belongs to is known, one individual MNL model per segment can be used. In contrast, if the assignment of customers to segments is unknown, the customers need to get probabilistically assigned to different segments implying that the customer segments become linked. Therefore, in the latter case the probability q_l of the membership for each customer segment and the MNL parameters β_l for all segments need to be jointly estimated, see Strauss et al. (2018).

The latent class multinomial logit (LC-MNL) is a special case of the MMNL under which the random MNL parameters follow a discrete distribution. This LC-MNL model is more convenient than the MMNL as the choice probabilities can be obtained in closed form. More formally, the deterministic part of the utility of product i for customer segment l is defined by $u_{il} = \beta_l^T z_i$ and the probability of choosing product i is given by

$$p_i(S) = \sum_{l \in L} q_l \cdot \frac{e^{u_{il}}}{1 + \sum_{j \in S} e^{u_{jl}}}.$$

In general, the MMNL model is able to approximate the choice probabilities of any choice model within the RUM framework arbitrarily close under mild regularity conditions (McFadden and Train 2000). Hence, this model provides a more substantial power in capturing customer choice behaviour compared to the MNL model. However, this superiority comes at the price of increasing computational complexity (Strauss et al. 2018).

A.1.3 Nested logit

Under the nested logit (NL) model, it is assumed that the choice set can be partitioned into K disjoint subsets called nests (Heiss 2002) in a way such that the IIA property holds within each nest but not across different nests (Strauss et al. 2018). The NL model can be identified as member of the RUM framework by assuming that the random components ϵ_i follow a general extreme value distribution that allows the alternatives within a nest to have mutually correlated error terms (Heiss 2002). To be precise, the values of the measure τ_k of the mutual correlation of the error terms of all alternatives within the nest k must lie in the unit interval (Heiss 2002).

Under the nested logit model, the probability $p_i(S)$ of purchasing a product i can be decomposed in two parts—the probability of choosing an alternative from the nest $k(i)$ to which product i belongs and the conditional probability to choose exactly alternative i given that some alternative of the nest $k(i)$ to which product i belongs is chosen (Heiss 2002).

- The probability that product i is purchased given that some alternative in its nest $k(i)$ is chosen is given by

$$\frac{e^{u_i/\tau_{k(i)}}}{\sum_{j\in k(i)} e^{u_j/\tau_{k(i)}}},$$

  where τ_k represents a measure of the mutual correlation of the error terms of all alternatives within the nest k, i.e. in the above formula of the nest $k(i)$ to which product i belongs.
- The probability that the customer purchases a product from nest $k(i)$ is obtained by

$$\frac{e^{\tau_{k(i)}\cdot\tilde{u}_{k(i)}}}{\sum_{k=1}^{K} e^{\tau_k\tilde{u}_k}},$$

  where $\tilde{u}_k = \ln(\sum_{j\in k} e^{u_j/\tau_k})$ represents the expected value of the utility an individual obtains from the alternatives in nest k.

Finally, the probability for choosing product i is calculated as the product of the probability that the customer purchases a product from nest $k(i)$ and the probability that product i is purchased given that some alternative of the nest $k(i)$ to which product i belongs is purchased, i.e.

$$p_i(S) = \frac{e^{u_i/\tau_{k(i)}}}{\sum_{j\in k(i)} e^{u_j/\tau_{k(i)}}} \cdot \frac{e^{\tau_{k(i)}\cdot\tilde{u}_{k(i)}}}{\sum_{k=1}^{K} e^{\tau_k\tilde{u}_k}} = \frac{e^{u_i/\tau_{k(i)}}}{e^{\tilde{u}_{k(i)}}} \cdot \frac{e^{\tau_{k(i)}\tilde{u}_{k(i)}}}{\sum_{k=1}^{K} e^{\tau_k\tilde{u}_k}}.$$

A.1.4 Paired combinatorial logit

Another choice model that is compatible with the RUM framework is the paired combinatorial logit (PCL) model. The PCL model allows for correlations between the utilities of any pair of products to capture situations where the preference of a customer for a particular product offers insights into the customer's attitude towards another product.

Under the PCL model, all products are grouped into nests of size two such that the collection of nests is represented by the set $M = \{(i,j)|i \neq j; i,j \in N\}$ of ordered pairs. For each nest (i,j), its dissimilarity parameter $\gamma_{ij} \in [0,1]$ characterizes the correlation between the utilities of products i and j. Based on this, the choice process for each arriving customer under the PCL can be modelled in two stages. In the first stage, the customer picks one of the $n(n-1)$ nests or leaves without a purchase. The preference weight for nest (i,j) is $V_{ij}(S)^{\gamma_{ij}}$ with $V_{ij}(S) = v_i^{1/\gamma_{ij}} \cdot 1_{\{i \in S\}} + v_j^{1/\gamma_{ij}} \cdot 1_{\{j \in S\}}$, where $v_i = e^{u_i} \geq 0$ denotes the preference weight for $i \in N \cup \{0\}$. Then, the probability that a customer picks nest (i,j) given that assortment S is offered is given by

$$P_{ij}(S) = \frac{V_{ij}(S)^{\gamma_{ij}}}{v_0 + \sum_{(k,l) \in M} V_{kl}(S)^{\gamma_{kl}}}.$$

Second, if the customer decides to make a purchase in nest (i,j), product i is chosen with probability

$$P_{i|ij}(S) = \frac{v_i^{1/\gamma_{ij}} \cdot 1_{\{i \in S\}}}{V_{ij}(S)}.$$

Finally, the probability that product i is chosen is obtained via

$$p_i(S) = \sum_{j \in N : j \neq i} P_{i|ij}(S) P_{ij}(S).$$

As discussed in Koppelman and Wen (2000), the estimation of the parameters of the PCL model is advantageous over the NL model since there is no need to search among numerous NL nesting structures. Moreover, the authors provide empirical evidence that the PCL model is indeed statistically superior to the MNL and NL models.

A.1.5 Exponomial

In contrast with the MNL or the NL model where the customers' willingness to pay distribution is assumed to be positively skewed, the exponomial choice

model (EXP) proposed by Alptekinoğlu and Semple (2016) assumes a negatively skewed distribution of customer utilities. This model is particularly suitable for situations in which the customer is well informed about products and their values such that his willingness to pay distribution is negatively skewed because he would be deterred by the prospect of overpaying (see Alptekinoğlu and Semple 2016).

The exponomial choice model can be identified as member of the RUM framework by assuming that the random components ϵ_i follow a negative exponential distribution with rate λ, hence the name 'exponomial' choice model. Under the assumption that the deterministic component u_i of the utility is sorted increasingly, i.e. $u_1 \leq u_2 \leq ... \leq u_n$, the probability to select a product i from the offer set S is given by

$$p_i(S) = \frac{\exp\left(-\lambda \sum_{j=i}^{n}(u_j - u_i)\right)}{n-i+1} - \sum_{k=1}^{i-1} \frac{\exp\left(-\lambda \sum_{j=k}^{n}(u_j - u_k)\right)}{(n-k)(n-k+1)}.$$

These probabilities can be obtained in closed form as the loglikelihood function is concave, and thus, maximum likelihood estimation can be used to determine the model parameters defining the deterministic component of the utility. Note that without loss of generality, one typically takes $\lambda = 1$ and rescales the u's accordingly.

A.1.6 Markov chain choice model

Under the Markov chain choice model (MCC) proposed by Blanchet et al. (2016), the consumer choice process is represented by a Markov chain with $N + 1$ states. Each state i corresponds to a product or the no-purchase option. If product i is offered, a customer that arrives at state i purchases this product; otherwise the customer proceeds to another state j. The arrival probability of state i is denoted by v_i and can be interpreted as the probability that product $i \in N \cup \{0\}$ is selected by the customer. The transition probability from state i to state j is denoted by ρ_{ij} and can be interpreted as the probability of substituting product i with product j in case it is unavailable. Every state is connected with state 0 that represents the no-purchase option. This implies that the customer can decide to not purchase anything at any time (see Strauss et al. 2018).

The model is fully defined by the parameter vectors $\boldsymbol{v}$ and $\boldsymbol{\rho}$. These parameters can be estimated using the expectation–maximization algorithm proposed by Şimşek and Topaloglu (2018). Moreover, the MCC belongs to the class of random utility-based choice models (Berbeglia 2016) and can approximate any choice model within the RUM framework under mild assumptions (Blanchet et al. 2016). Therefore, this choice model can be used as approximate model when the true underlying choice model is known but the corresponding assortment optimization problem is known to be NP-hard as e.g. in case of the MMNL and the NL model. According to Blanchet et al. (2016), the derived solution

under the Markov chain choice model is near-optimal in case the MCC is a good approximation to the true underlying model.

A.2 Nonparametric choice models

As mentioned before, the previously introduced parametric choice models fully depend on the choice of their underlying parameters which are typically unknown and need to be chosen or estimated. In addition, assumptions must be made about the relevant model covariates and about the functional form of the relationship between product attributes, utility values and choice probabilities, see e.g. Jagabathula and Rusmevichientong (2016) and Berbeglia et al. (2021a). Yet, the specified assumptions may not adequately capture the actual choice behaviour (Strauss et al. 2018).

In contrast, nonparametric choice models are not built upon any assumption on the data structure but are solely shaped by data and thus by design do not suffer from this problem. On the other hand, nonparametric choice models typically do not allow for extrapolation and prediction of changes in the demand pattern due to changes in a product attribute since such nonparametric models are typically designed as ranked lists of preferences, also referred to as customer types, see Berbeglia et al. (2021a). Under such rank list-based models, the customer chooses the highest-ranking offered product or leaves without making a purchase in case none of the offered products ranks higher than the no-purchase option. Demand is modelled by a probability distribution over all customer types, see e.g. Jagabathula and Rusmevichientong (2016).

Note that the potential number of customer types is factorial though the actual number of underlying customer types might be way smaller (Jena et al. 2020). Overall, this model is quite general and subsumes various choice models typically considered in assortment optimization such as the MNL (Mahajan and van Ryzin 2001).

Appendix B: Related areas

All studies mentioned so far solely consider assortment problems where the customer is offered a selection of products and decides whether and which product(s) to purchase. However, there also exist settings with two-sided markets as it is e.g. the case for matching platforms. This setting is referred to as two-sided assortment optimization. We briefly introduce it in Section B.1. Moreover, as mentioned before, retailers often not only face decisions regarding the selection of products to offer but additionally face further tasks such as to determine the prices for the offered SKUs, the number of units to stock per SKU or the shelf or display space allocation. This section is targeted to provide an overview of these settings. To be precise, we introduce the joint assortment and price optimization problem in Section B.2, the joint assortment and inventory level optimization in Section B.3, and the joint assortment optimization and shelf-space allocation task in Section B.4. Recently, researchers

also consider further joint assortment optimization settings which we briefly introduce in Section B.5. Finally, in Section B.6, we briefly introduce the research area of facility location planning which is somehow similar to assortment optimization.

B.1 Two-sided assortment optimization

The two-sided assortment optimization is targeted to extend the literature on classic one-sided decision making to two-sided markets by considering the effect of choice decisions by both sides on the final outcome. This setting typically occurs on two-sided matching platforms, where the platform has a set of suppliers and consumers each of which has a utility associated with the opposite side as well as a utility for the outside option. The platform offers a set of suppliers to each consumer and a set of consumers to each supplier. All participants independently select at most one individual from the assortment offered to them. A supplier-consumer match occurs when both sides select each other. The selection probabilities are typically determined by well-known choice models. Each successful match generates a certain revenue for the platform; unsuccessful matches do not generate any revenue. The overall objective is to choose an assortment family that maximizes the expected revenue for the matching. The supplier and consumer selection processes can proceed either sequential or simultaneous.

For example Torrico et al. (2021) consider the assortment problem under a two-sided sequential matching process and provide constant-factor guarantees for the general case as well as for the extension to cardinality constraints.

Likewise, Ashlagi et al. (2022) study the two-sided matching problem between customers and suppliers where the platform offers a menu of suppliers to each customer and the customers choose simultaneously and independently to either select a supplier from their menu or remain unmatched. Suppliers then see the set of customers that have selected them and choose to either match with one of these customers or remain unmatched. The authors show that this problem is strongly NP-hard and provide an efficient algorithm that achieves a constant-factor approximation guarantee for the optimal expected number of matches.

In contrast with the previously mentioned studies, Ahmed et al. (2022) consider the two-sided assortment optimization while assuming that the matching process takes places simultaneously. The authors assume that the selection probabilities for both sides follow MNL choice models and prove that this problem is NP-hard even when the number of suppliers is limited to two. The authors propose a mixed-integer linear programming formulation and develop relaxations that provide upper and lower bounds whose practical utility is demonstrated on synthetic data.

Similarly, Rios et al. (2022) study the two-sided assortment problem within a dating platform setting where the set of potential partners to be shown to each user should be dynamically selected per time period in order to maximize the expected number of matches. The authors model this task as dynamic optimization problem and propose a family of heuristics for its solution. For further information on two-sided assortment optimization, we refer the interested reader to the above mentioned literature and the references therein.

B.2 Joint assortment optimization and pricing

In all studies reviewed until now, the product prices are assumed to be exogenously given. However, in practice this assumption only holds when the retailer adopts the Manufacturer Suggested Retail Price (MSRP) or a function thereof such as a 10% markup for all products. In most other cases, the retail prices for all products contained in the assortment need to be set by the retailer, which is referred to as pricing. It is natural to select optimal assortment and prices at once. This task is referred to as joint assortment optimization and pricing and is extensively studied in the literature.

For example Jagabathula and Rusmevichientong (2016) propose a nonparametric framework to joint assortment optimization and pricing where each customer is represented by a preference list over all alternatives and a price threshold. The customer follows a two-stage choice process by first considering the set of products with prices less than their threshold value and subsequently choosing the most preferred product from the remaining consideration set. The authors propose a tractable expectation maximization framework for model fitting along with an efficient algorithm to determine the profit-maximizing combination of offer set and price.

Miao and Chao (2020) consider the dynamic joint assortment optimization and pricing problem under the MNL model with sequentially arriving customers when the firm has limited prior knowledge about the consumer demand. The authors design a learning algorithm balancing the trade-off between demand learning and revenue extraction, evaluate the algorithms performance using Bayesian regret and provide an instance-independent upper bound for the Bayesian regret of the algorithm. Numerical experiments provide empirical evidence for the practical applicability of the proposed approach.

Likewise, Gao (2021) study joint assortment optimization and pricing problems under a variant of the MNL model with impatient customers where the customer sequentially views the assortment of available products. The maximum number of viewed stages is determined by the customers patience level. The authors provide a 87.8% approximation algorithm for this problem.

Chen et al. (2021b) consider joint pricing and assortment decisions under a logit model-based framework that takes customer features into account. This model provides a significant advantage when insufficient data for every customer are available. For further details on joint assortment optimization and pricing, we refer the interested reader to the above mentioned literature and the references therein.

Finally, another research area related to the classic assortment optimization is the product line design. To tackle the task of product line design, typically a two-step approach is followed. First, a set of candidate products is determined. In the second step, a product line selection—i.e. assortment optimization—and pricing problem is solved on the set of candidate products to determine the optimal product line. Schön (2010a) proposes an exact approach for determining the profit-maximizing product line under the consideration of continuous prices when customers are assumed to choose according to an attraction choice model.

Likewise, Schön (2010b) studies an approach to find the optimal, profit-maximizing product line under a personalized or group pricing strategy in markets with multiple heterogeneous costumers.

B.3 Joint assortment and inventory level optimization

Besides pricing, the retailer also needs to determine how many items to stock for each SKU. This problem is particularly relevant in offline channels as brick-and-mortar stores typically have only limited capacity both on their shelves and in the warehouse but need to prevent stock-outs. It is only natural to optimize offer set and corresponding inventory levels at the same time. This is referred to as joint assortment and inventory level optimization and frequently studied in the literature.

Transchel et al. (2022) consider the joint assortment and inventory planning problem for vertically differentiated products under consumer-driven substitution where the demand for each product and the stockout-based substitution rates are derived from a customer's utility function and a random market size. The authors propose a two-step integral solution approach where first the initial purchasing probabilities of all products are determined along with the substitution matrices of all possible product-availability combinations. Next, the inventory levels are obtained by iteratively solving a sequence of two-product problems. The authors provide evidence that joint assortment and inventory planning while considering stockout-based substitution is particularly essential when both the profit margin and demand uncertainty are high.

Honhon et al. (2010) consider the single-period assortment and inventory problem under stockout-based substitution where customers can be assigned to different customer types. Each customer type corresponds to a preference ordering among products and purchases the highest-ranked offered product if any. The authors solve the optimal assortment problem using a dynamic programming formulation and establish structural properties of the optima of the value function that enable the solvability of the problem in pseudopolynomial time. Moreover, the authors give a heuristic for the case when the proportion of customers of each type is random and provide empirical evidence of the applicability of their approach.

Moreover, Martínez-de Albéniz and Kunnumkal (2022) consider the joint assortment and inventory planning problem under stockout-based substitution in a setting with inventory replenishment. The authors develop an accurate approximation for the multi-product case by making use of the existing closed-form solution for the single-product case.

Likewise, Aouad et al. (2018b) study the joint assortment and inventory planning problem under the MNL model where stock-out events cause dynamic substitution effects. The authors propose an algorithm for dynamic assortment planning under the MNL model and derive a constant-factor guarantee for a broad class of demand distributions.

Finally, Katsifou et al. (2014) consider the joint assortment, inventory level, and pricing problem under the MNL model in the setting where a retailer carries a product assortment consisting of both standard products and short-lived special products. This setting is intended to increase store traffic by attracting heterogeneous classes

of customers, which in turn increases the sales of standard products due to potential cross-selling effects as customers who are primarily attracted by special products might also buy standard products. The authors propose an optimization model and an iterative heuristic and provide empirical evidence that retailers might benefit from carrying low-priced special products on top of the standard product assortment. For further information on joint assortment and inventory level optimization, we refer the interested reader to the above introduced literature and the references therein.

B.4 Joint assortment optimization and shelf-space allocation

Shelf-space allocation refers to the task of allocating the available shelf space among all products included in the assortment. To this end, one typically defines facing quantities to individual products while restricting the available shelf space. One common assumption in shelf-space optimization is that the sales of a product are directly linked to the space allocated to it and the space allocated to its competitors. Hence, instead of using choice models to capture consumer demand, researchers rather apply so-called space-elastic functions to estimate the demand when it comes to shelf-space allocation tasks. Space elasticity refers to the sensitivity of the customer to the inventory (or number of facings) displayed in terms of quantity bought (see Hariga et al. 2007). It is again only natural to jointly optimize the offered assortment along with the shelf-space allocation. This setting is extensively studied in the literature.

For example, Hübner and Schaal (2017) consider the joint assortment optimization and shelf-space allocation problem. The authors formulate a model that maximizes the retailer's profit by selecting the optimal assortment and assigning limited shelf-space to items while considering stochastic and space-elastic demand as well as out-of-assortment and out-of-stock substitution effects. The authors develop a specialized heuristic that yields near-optimal results even for large-scale problems and find that space elasticity and substitution effects have a significant impact on profits, assortment size and facing decisions.

Hübner et al. (2020) consider the joint assortment and shelf-space allocation problem for two-dimensional, tilted shelves as e.g. used for cheese or clothes. The authors develop a decision model that optimizes the assortment selection and the assignment of items to a space-restricted, tilted shelf while accounting for stochastic demand, space elasticity, and substitution effects. To solve this problem, a specialized heuristic based on genetic algorithms that yields near-optimal results is proposed. For further information on joint assortment optimization and shelf-space allocation, we refer the interested reader to the above introduced literature and the references therein as well as to Hübner and Kuhn (2012) for an excellent comprehensive review.

Obviously, the shelf-space allocation problem is only relevant in offline settings. However, it can easily be transferred to the online setting by assuming that a retailer's website is of limited space and it needs to be determined which product is offered at which position on the website. The joint optimization of assortment and

product display position is exemplary tackled by Chen and Jiang (2020a), who study the joint assortment and display position problem under the nested logit model. The authors formulate this problem as a nonlinear binary integer programming model, develop a dynamic programming-based solution approach for obtaining optimal assortment-position assignments, and find that it is not necessarily better to put the most attractive products in the best positions. Finally, the authors discuss the extension to the joint assortment-position-price optimization problem.

B.5 Further joint assortment optimization problems

Besides the popular joint assortment problems introduced so far, researchers additionally consider a variety of other problem settings. For example Wang et al. (2022a) consider a joint advertising and assortment optimization problem under the MNL model, where the objective is to find an optimal product assortment and optimal advertising strategies for them. To be precise, the authors assume that the preference weight of a product can be increased by advertising it. The degree of improvement is determined by the effectiveness of the advertisement and the amount of advertising efforts allocated to that product. The authors show that revenue-ordered assortments remain optimal in the uncapacited case and provide a relaxation to efficiently obtain near-optimal solutions under cardinality constraints.

Wang et al. (2023e) consider the joint optimization of offline and online decisions. The former comprises decisions regarding the product-design characteristics such as price, capacity or return eligibility; the latter involves the dynamic assortment optimization over a selling season. This setting e.g. occurs in practice when determining product discounts or a products' return eligibility. The authors formulate an optimization problem combining the impact of both offline and online decisions on the expected revenue. To determine the product design, the authors reformulate a choice-based deterministic linear program, solve its continuous relaxation, and round the resulting solution. A dynamic assortment policy achieving at least a constant fraction of the expected revenue of the choice-based deterministic linear program is obtained using value function approximations. Combining both results, the authors provide an approximate solution with performance guarantees to the joint product design and assortment optimization.

Haase and Müller (2020) study the joint assortment and product design optimization problem under the mixed logit model with deterministic customer segments. In this setting, the objective is to select an offerset of predetermined size and decide on the attributes of each product such that a function of market share is maximized. The authors develop a mixed-integer nonlinear program (MINLP) that is solved by generic solvers and provide empirical evidence that even large instances can be solved in reasonable time. Similarly, Jiao and Zhang (2005) study the product portfolio planning problem, i.e. the task of selecting an optimal mix of products and attributes to be offered. The authors consider a shared-surplus maximization model that takes customer preferences and choice behaviour as well as platform-based product costing into account and propose a stochastic mixed-integer nonlinear program while jointly considering customer concerns and operational implications.

Their approach is later on adapted and enhanced by Müller and Haase (2016) who propose some changes such as demand model calibration, deterministic customer surplus, and an effective objective function.

El Housni and Topaloglu (2022) consider a joint assortment optimization and customization problem under the MMNL model. In this setting, a firm first selects an assortment of products to carry subject to a cardinality constraint. When a customer of certain type arrives, the pre-selected assortment is adjusted to the observed customer type by possibly dropping certain products from the assortment. The overall goal under this setting is to determine the assortment to carry as well as the customized assortments per customer type. Such settings exemplary arise in online platforms where retailers commit to a selection of products before the start of the selling season and adjust the the displayed assortment per customer type. The authors show NP-hardness of this problem and provide an approximation framework with performance guarantee.

Chen et al. (2023b) study the 'recommendation at checkout' problem, where each arriving customer type is defined by a primary item of interest that this customer might add to his shopping card. In case the item is added to the card, the retailer aims at recommending an assortment of add-ons to the customer that go along with the primary item. The authors derive an algorithm with 1/4-competitive ratio guarantee under adversarial arrivals.

Another research area related to assortment optimization is the so-called bundling. Bundling addresses the question which products should be combined to product bundles to be offered together. We distinguish two variants of bundling—pure and mixed bundling. The former occurs when the products contained in a bundle can only be purchased together but not on their own. In contrast, mixed bundling refers to the case when customers can decide whether to purchase the entire bundle or its individual components. The retailer needs to jointly determine which products to bundle and offer to the customer and the prices demanded for the product bundles. Ettl et al. (2019) address this topic by constructing a model that recommends personalized discounted product bundles to online shoppers while considering the trade-off between profit maximization and inventory management. The authors determine analytical performance guarantees illustrating the complexity of this joint bundling, assortment and pricing problem and provide empirical evidence of the applicability of their approach.

B.6 Facility location planning

Another research area that appears to be similar to assortment optimization is the topic of facility location planning. Instead of determining the optimal selection of products to be offered to a customer, facility location planning deals with the task of determining the optimal locations of facilities. Such facilities can be companies, (manufacturing) plants, warehouses, or facilities such as schools, emergency services, fire stations and much more (see Domschke and Krispin 1997).

Like in assortment optimization, the applications of facility planning are manifold, resulting in an ever growing family of location allocation models—ranging

from simple linear, single-stage, single-product, uncapacited, deterministic models to nonlinear probabilistic models (see Klose and Drexl 2005). Particularly, there are a growing number of articles considering the MNL model in facility location models (e.g. Aros-Vera et al. 2013; Benati and Hansen 2002; Haase and Müller 2013, 2014). Likewise, a broad range of algorithms including local search and mathematical programming-based approaches for tackling this problem are proposed (see Klose and Drexl 2005). For example, only considering locational decisions, there exist different linear formulations of the MNL yielding mixed-integer linear programs that are compared in Haase and Müller (2014).

For more information, we refer the interested reader to Haase et al. (2019) who provide an overview of publications on facility location planning in the public sector or to Melo et al. (2009) who provide a review of facility location models in the context of supply chain management.

Funding Open Access funding enabled and organized by Projekt DEAL.

Declarations

Conflict of interest The authors have no relevant financial or non-financial interests to disclose.

References

Agrawal S, Avadhanula V, Goyal V, Zeevi A (2019) MNL-bandit: a dynamic learning approach to assortment selection. Oper Res 67(5):1453–1485. https://doi.org/10.1287/opre.2018.1832

Aguirregabiria V, Mira P (2010) Dynamic discrete choice structural models: a survey. J Econom 156(1):38–67. https://doi.org/10.1016/j.jeconom.2009.09.007

Ahmed A, Sohoni M, Bandi C (2022) Parameterized approximations for the two-sided assortment optimization. Oper Res Lett 50(4):399–406. https://doi.org/10.1016/j.orl.2022.04.002

Alfandari L, Hassanzadeh A, Ljubić I (2021) An exact method for assortment optimization under the nested logit model. Eur J Oper Res 291(3):830–845. https://doi.org/10.1016/j.ejor.2020.12.007

Alptekinoğlu A, Semple JH (2016) The exponomial choice model: a new alternative for assortment and price optimization. Oper Res 64(1):79–93. https://doi.org/10.1287/opre.2015.1459

Alptekinoğlu A, Semple JH (2021) Heteroscedastic exponomial choice. Oper Res 69(3):841–858. https://doi.org/10.1287/opre.2020.2074

Aouad A, Levi R, Segev D (2018a) Approximation algorithms for dynamic assortment optimization models. Math Oper Res 44(2):487–511. https://doi.org/10.1287/moor.2018.0933

Aouad A, Levi R, Segev D (2018b) Greedy-like algorithms for dynamic assortment planning under multinomial logit preferences. Oper Res 66(5):1321–1345. https://doi.org/10.1287/opre.2018.1734

Aouad A, Farias V, Levi R (2020) Assortment optimization under consider-then-choose choice models. Manag Sci 67(6):3368–3386. https://doi.org/10.1287/mnsc.2020.3681

Aouad A, Feldman J, Segev D, Zhang D (2021) The click-based MNL model: a novel framework for modeling click data in assortment optimization. SSRN. https://doi.org/10.2139/ssrn.3340620
Aouad A, Feldman J, Segev D (2022) The exponomial choice model for assortment optimization: an alternative to the MNL model? Manag Sci 69(5):2814–2832. https://doi.org/10.1287/mnsc.2022.4492
Aros-Vera F, Marianov V, Mitchell JE (2013) p-Hub approach for the optimal park-and-ride facility location problem. Eur J Oper Res 226(2):277–285. https://doi.org/10.1016/j.ejor.2012.11.006
Aouad A, Désir A (2022) Representing random utility choice models with neural networks. arXiv. https://doi.org/10.48550/arXiv.2207.12877
Ashlagi I, Krishnaswamy AK, Makhijani R, Saban D, Shiragur K (2022) Technical note—assortment planning for two-sided sequential matching markets. Oper Res 70(5):2784–2803. https://doi.org/10.1287/opre.2022.2327
Atamtürk A, Gómez A (2020) Submodularity in conic quadratic mixed 0–1 optimization. Oper Res 68(2):609–630. https://doi.org/10.1287/opre.2019.1888
Bai Y, Feldman J, Segev D, Topaloglu H, Wagner L (2023a) Assortment optimization under the multi-purchase multinomial logit choice model. Oper Res Articles Adv. https://doi.org/10.1287/opre.2023.2463
Bai Y, Feldman J, Topaloglu H, Wagner L (2023b) Assortment optimization under the multinomial logit model with utility-based rank cutoffs. Oper Res Articles Adv. https://doi.org/10.1287/opre.2021.0060
Bechler G, Steinhardt C, Mackert J (2021) On the linear integration of attraction choice models in business optimization problems. Oper Res Forum 2:12. https://doi.org/10.1007/s43069-021-00056-1
Ben-Akiva ME, Lerman SR (1985) Discrete choice analysis: theory and application to travel demand. MIT press series in transportation studies, MIT Press, Cambridge
Ben-Akiva M et al (1994) Combining revealed and stated preferences data. Mark Lett 5:335–349. https://doi.org/10.1007/BF00999209
Benati S, Hansen P (2002) The maximum capture problem with random utilities: problem formulation and algorithms. Eur J Oper Res 143(3):518–530. https://doi.org/10.1016/S0377-2217(01)00340-X
Berbeglia G (2016) Discrete choice models based on random walks. Oper Res Lett 44(2):234–237. https://doi.org/10.1016/j.orl.2016.01.009
Berbeglia G, Garassino A, Vulcano G (2021a) A comparative empirical study of discrete choice models in retail operations. Manage Sci 68(6):4005–4023. https://doi.org/10.1287/mnsc.2021.4069
Berbeglia G, Flores A, Gallego G (2021b) Refined assortment optimization. SSRN. https://doi.org/10.2139/ssrn.3778413
Bernstein F, Kök AG, Xie L (2015) Dynamic assortment customization with limited inventories. Manuf Serv Oper Manag 17(4):538–553. https://doi.org/10.1287/msom.2015.0544
Bernstein F, Modaresi S, Sauré D (2019) A dynamic clustering approach to data-driven assortment personalization. Manag Sci 65(5):2095–2115. https://doi.org/10.1287/mnsc.2018.3031
Bernstein F, Modaresi S, Sauré D (2022) Exploration optimization for dynamic assortment personalization under linear preferences. SSRN. https://doi.org/10.2139/ssrn.4115721
Bertsimas D, Mišić V (2015) Data-driven assortment optimization. Working paper, MIT Sloan School of Management, Cambridge, MA
Besbes O, Gur Y, Zeevi A (2015) Optimization in online content recommendation services: beyond click-through rates. Manuf Serv Oper Manag 18(1):15–33. https://doi.org/10.1287/msom.2015.0548
Blanchet J, Gallego G, Goyal V (2016) A Markov chain approximation to choice modeling. Oper Res 64(4):886–905. https://doi.org/10.1287/opre.2016.1505
Cachon GP, Terwiesch C, Xu Y (2005) Retail assortment planning in the presence of consumer search. Manuf Serv Oper Manag 7(4):330–346. https://doi.org/10.1287/msom.1050.0088
Cai Z, Wang H, Talluri K, Li X (2022) Deep learning for choice modeling. arXiv. https://doi.org/10.48550/arXiv.2208.09325
Campbell BM (1969) The existence of evoked set and determinants of its magnitude in brand choice behavior. Ph.D. thesis, Columbia University
Cao Y, Rusmevichientong P, Topaloglu H (2022) Revenue management under a mixture of independent demand and multinomial logit models. Oper Res 71(2):603–625. https://doi.org/10.1287/opre.2022.2333
Caro F, Gallien J (2007) Dynamic assortment with demand learning for seasonal consumer goods. Manag Sci 53(2):276–292. https://doi.org/10.1287/mnsc.1060.0613

Caro F, Martínez-de Albéniz V, Rusmevichientong P (2014) The assortment packing problem: multiperiod assortment planning for short-lived products. Manag Sci 60(11):2701–2721. https://doi.org/10.1287/mnsc.2014.1991

Chen Y-C, Mišić V (2021) Assortment optimization under the decision forest model. SSRN. https://doi.org/10.2139/ssrn.3812654

Chen R, Jiang H (2020a) Assortment optimization with position effects under the nested logit model. Nav Res Logist 67(1):21–33. https://doi.org/10.1002/nav.21879

Chen R, Jiang H (2020b) Capacitated assortment and price optimization under the nested logit model. J Global Optim 7:895–918. https://doi.org/10.1007/s10898-020-00896-x

Chen X, Shi C, Wang Y, Zhou Y (2021a) Dynamic assortment planning under nested logit model. Prod Oper Manag 30(1):85–102. https://doi.org/10.1111/poms.13258

Chen X, Owen Z, Pixton C, Simchi-Levi D (2021b) A statistical learning approach to personalization in revenue management. Manag Sci 68(3):1923–1937. https://doi.org/10.1287/mnsc.2020.3772

Chen X, Li J, Li M, Zhao T, Zhou Y (2022) Assortment optimization under the multivariate MNL model. arXiv. https://doi.org/10.48550/arXiv.2209.15220

Chen X, Krishnamurthy A, Wang Y (2023a) Robust dynamic assortment optimization in the presence of outlier customers. Oper Res Articles Adv. https://doi.org/10.1287/opre.2020.0281

Chen X, Ma W, Simchi-Levi D, Xin L (2023b) Assortment planning for recommendations at checkout under inventory constraints. Math Oper Res Articles Adv. https://doi.org/10.1287/moor.2023.1357

Cheung WC, Simchi-Levi D (2017) Thompson sampling for online personalized assortment optimization problems with multinomial logit choice models. SSRN. https://doi.org/10.2139/ssrn.3075658

Davis J, Gallego G, Topaloglu H (2014) Assortment optimization under variants of the nested logit model. Oper Res 62(2):250–273. https://doi.org/10.1287/opre.2014.1256

Davis J, Gallego G, Topaloglu H (2013) Assortment planning under the multinomial logit model with totally unimodular constraint structures. Working paper, Cornell University, Ithaca, NY

Debreu G (1960) Review of Individual choice behavior: a theoretical analysis by R.D. Luce. Am Econ Rev 50(1):186–188

Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm. J R Stat Soc: Ser B (Methodol) 39(1):1–22. https://doi.org/10.1111/j.2517-6161.1977.tb01600.x

Désir A, Goyal V, Segev D, Ye C (2019) Constrained assortment optimization under the Markov chain-based choice model. Manag Sci 66(2):698–721. https://doi.org/10.1287/mnsc.2018.3230

Désir A, Goyal V, Jagabathula S, Segev D (2021) Mallows-smoothed distribution over rankings approach for modeling choice. Oper Res 69(4):1206–1227. https://doi.org/10.1287/opre.2020.2085

Désir A, Goyal V, Jiang B, Xie T, Zhang J (2023) Robust assortment optimization under the Markov chain choice model. Oper Res Articles Adv. https://doi.org/10.1287/opre.2022.2420

Domschke W, Krispin G (1997) Location and layout planning. OR Spectrum 19:181–194. https://doi.org/10.1007/bf01545586

Dong J, et al. (2023) PASTA: Pessimistic assortment optimization. arXiv. https://doi.org/10.48550/arXiv.2302.03821

Doudchenko N, Drynkin E (2020) Estimation of discrete choice models: a machine learning approach. arXiv. https://doi.org/10.48550/arXiv.2010.08016

Durach CF, Kembro J, Wieland A (2017) A new paradigm for systematic literature reviews in supply chain management. J Supply Chain Manag 53(4):67–85. https://doi.org/10.1111/jscm.12145

El Housni O, Topaloglu H (2022) Joint assortment optimization and customization under a mixture of multinomial logit models: on the value of personalized assortments. Oper Res 71(4):1197–1215. https://doi.org/10.1287/opre.2022.2384

Ettl M, Harsha P, Papush A, Perakis G (2019) A data-driven approach to personalized bundle pricing and recommendation. Manuf Serv Oper Manag 22(3):461–480. https://doi.org/10.1287/msom.2018.0756

Feldman J (2017) Technical note: space constrained assortment optimization under the paired combinatorial logit model. SSRN. https://doi.org/10.2139/ssrn.3013321

Feldman J, Paul A (2019) Relating the approximability of the fixed cost and space constrained assortment problems. Prod Oper Manag 28(5):1238–1255. https://doi.org/10.1111/poms.12983

Feldman J, Segev D (2022) Technical note—the multinomial logit model with sequential offerings: algorithmic frameworks for product recommendation displays. Oper Res 70(4):2162–2184. https://doi.org/10.1287/opre.2021.2218

Feldman J, Topaloglu H (2015a) Bounding optimal expected revenues for assortment optimization under mixtures of multinomial logits. Prod Oper Manag 24(10):1598–1620. https://doi.org/10.1111/poms.12365
Feldman J, Topaloglu H (2015b) Capacity constraints across nests in assortment optimization under the nested logit model. Oper Res 63(4):812–822. https://doi.org/10.1287/opre.2015.1383
Feldman J, Topaloglu H (2017a) Revenue management under the Markov chain choice model. Oper Res 65(5):1322–1342. https://doi.org/10.1287/opre.2017.1628
Feldman J, Topaloglu H (2017b) Technical note—capacitated assortment optimization under the multinomial logit model with nested consideration sets. Oper Res 66(2):380–391. https://doi.org/10.1287/opre.2017.1672
Feldman J, Paul A, Topaloglu H (2019) Technical note—assortment optimization with small consideration sets. Oper Res 67(5):1283–1299. https://doi.org/10.1287/opre.2018.1803
Feldman J, Zhang D, Liu X, Zhang N (2021) Customer choice models vs. machine learning: finding optimal product displays on Alibaba. Oper Res 70(1):309–328. https://doi.org/10.1287/opre.2021.2158
Feng Q, Wang Z (2021) Dynamic multinomial logit choice model with network effect. SSRN. https://doi.org/10.2139/ssrn.3939717
Flores A, Berbeglia G, Van Hentenryck P (2019) Assortment optimization under the sequential multinomial logit model. Eur J Oper Res 273(3):1052–1064. https://doi.org/10.1016/j.ejor.2018.08.047
Gallego G, Topaloglu H (2014) Constrained assortment optimization for the nested logit model. Manag Sci 60(10):2583–2601. https://doi.org/10.1287/mnsc.2014.1931
Gallego G, Berbeglia G (2022) Bounds, heuristics, and prophet inequalities for assortment optimization. arXiv. https://doi.org/10.48550/arXiv.2109.14861
Gallego G, Iravani MM, Talebian M (2023) Constrained assortment optimization with satisficers consumers. SSRN. https://doi.org/10.2139/ssrn.4402473
Gallego G, Lu W (2021) An optimal greedy heuristic with minimal learning regret for the Markov chain choice model. SSRN. https://doi.org/10.2139/ssrn.3810470
Gallego G, Topaloglu H (2019) Revenue management and pricing analytics, Ch. 5. Springer, New York, pp 129–160
Gao P et al (2021) Assortment optimization and pricing under the multinomial logit model with impatient customers: sequential recommendation and selection. Oper Res 69(5):1509–1532. https://doi.org/10.1287/opre.2021.2127
Ghuge R, Kwon J, Nagarajan V, Sharma A (2021) Constrained assortment optimization under the paired combinatorial logit model. Oper Res 70(2):786–804. https://doi.org/10.1287/opre.2021.2188
Golrezaei N, Nazerzadeh H, Rusmevichientong P (2014) Real-time optimization of personalized assortments. Manag Sci 60(6):1532–1551. https://doi.org/10.1287/mnsc.2014.1939
Gong X-Y et al (2021) Online assortment optimization with reusable resources. Manag Sci 68(7):4772–4785. https://doi.org/10.1287/mnsc.2021.4134
Goutam K, Goyal V, Lam H (2020) Assortment optimization over dense universe is easy. SSRN. https://doi.org/10.2139/ssrn.3649233
Haase K, Müller S (2013) Management of school locations allowing for free school choice. Omega 41(5):847–855. https://doi.org/10.1016/j.omega.2012.10.008
Haase K, Müller S (2014) A comparison of linear reformulations for multinomial logit choice probabilities in facility location models. Eur J Oper Res 232(3):689–691. https://doi.org/10.1016/j.ejor.2013.08.009
Haase K, Knörr L, Krohn R, Müller S, Wagner M (2019) Facility location in the public sector. Springer, Cham, pp 745–764
Haase K, Müller S (2020) Constrained assortment optimization under the mixed logit model with design options. SSRN. https://doi.org/10.2139/ssrn.3624816
Han Y, Pereira F, Ben-Akiva M, Zegras C (2022) A neural-embedded discrete choice model: learning taste representation with strengthened interpretability. Transp Res Part B: Methodol 163:166–186. https://doi.org/10.1016/j.trb.2022.07.001
Hariga MA, Al-Ahmari A, Mohamed A-RA (2007) A joint optimisation model for inventory replenishment, product assortment, shelf space and display area allocation decisions. Eur J Oper Res 181(1):239–251. https://doi.org/10.1016/j.ejor.2006.06.025
Heiss F (2002) Structural choice analysis with nested logit models. Stata J 2(3):227–252. https://doi.org/10.1177/1536867X0200200301
Hensher DA, Rose JM, Greene WH (2005) Applied choice analysis: a primer. Cambridge University Press, Cambridge

Homer S, Selman AL (2011) Computability and complexity theory. Texts in computer science. Springer, New York

Honhon D, Gaur V, Seshadri S (2010) Assortment planning and inventory decisions under stockout-based substitution. Oper Res 58(5):1364–1379. https://doi.org/10.1287/opre.1090.0805

Honhon D, Jonnalagedda S, Pan XA (2012) Optimal algorithms for assortment selection under ranking-based consumer choice models. Manuf Serv Oper Manag 14(2):279–289. https://doi.org/10.1287/msom.1110.0365

Honhon D, Pan XA, Sreelata J (2020) In-out algorithm for assortment planning under a ranking-based consumer choice model. Oper Res Lett 48(3):309–316. https://doi.org/10.1016/j.orl.2020.03.005

Howard J, Sheth J (1969) The theory of buyer behavior. Wiley, New York

Hübner A, Kuhn H (2012) Retail category management: state-of-the-art review of quantitative research and software applications in assortment and shelf space management. Omega 40(2):199–209. https://doi.org/10.1016/j.omega.2011.05.008

Hübner A, Schaal K (2017) An integrated assortment and shelf-space optimization model with demand substitution and space-elasticity effects. Eur J Oper Res 261(1):302–316. https://doi.org/10.1016/j.ejor.2017.01.039

Hübner A, Schäfer F, Schaal K (2020) Maximizing profit via assortment and shelf-space optimization for two-dimensional shelves. Prod Oper Manag 29(3):547–570. https://doi.org/10.1111/poms.13111

Hu B, Jin Q, Long D (2022) Robust assortment revenue optimization and satisficing. SSRN. https://doi.org/10.2139/ssrn.4045001

Jagabathula S (2016) Assortment optimization under general choice. SSRN. https://doi.org/10.2139/ssrn.2512831

Jagabathula S, Rusmevichientong P (2016) A nonparametric joint assortment and price choice model. Manag Sci 63(9):3128–3145. https://doi.org/10.1287/mnsc.2016.2491

Jagabathula S, Mitrofanov D, Vulcano G (2023) Demand estimation under uncertain consideration sets. Oper Res Articles Adv. https://doi.org/10.1287/opre.2022.0006

Jena SD, Lodi A, Palmer H, Sole C (2020) A partially ranked choice model for large-scale data-driven assortment optimization. INFORMS J Optim 2(4):297–319. https://doi.org/10.1287/ijoo.2019.0037

Jiang S, Nip K-M (2022) An enhanced conic reformulation for capacity-constrained assortment optimization under the mixture of multinomial logit model. J Oper Res Soc China. https://doi.org/10.1007/s40305-022-00438-0

Jiao J, Zhang Y (2005) Product portfolio planning with customer-engineering interaction. IIE Trans 37(9):801–814. https://doi.org/10.1080/07408170590917011

Jin Q, Wang Q, Han Y (2023) Pricing and assortment optimization under logit based choice models with tree structured consideration sets. SSRN. https://doi.org/10.2139/ssrn.4129238

Kallus N, Udell M (2020) Dynamic assortment personalization in high dimensions. Oper Res 68(4):1020–1037. https://doi.org/10.1287/opre.2019.1948

Karampatsa M, Grigoroudis E, Matsatsinis NF (2017) Retail category management: a review on assortment and shelf-space planning models. In: Grigoroudis E, Doumpos M (eds) Operational research in business and economics. Springer, Cham, pp 35–67

Katsifou A, Seifert R, Tancrez J-S (2014) Joint product assortment, inventory and price optimization to attract loyal and non-loyal customers. Omega 46:36–50. https://doi.org/10.1016/j.omega.2014.02.002

Keane MP, Wolpin KI (2009) Empirical applications of discrete choice dynamic programming models. Rev Econ Dyn 12(1):1–22. https://doi.org/10.1016/j.red.2008.07.001

Klose A, Drexl A (2005) Facility location models for distribution system design. Eur J Oper Res 162(1):4–29. https://doi.org/10.1016/j.ejor.2003.10.031

Kök AG, Fisher ML, Vaidyanathan R (2008) Assortment planning: review of literature and industry practice. In: Agrawal N, Smith SA (eds) Retail supply chain management: quantitative models and empirical studies. Springer, Boston, pp 99–153

Koppelman F, Wen C-H (2000) The paired combinatorial logit model: properties, estimation and application. Transp Res Part B: Methodol 34(2):75–89. https://doi.org/10.1016/S0191-2615(99)00012-0

Kunnumkal S (2015) On upper bounds for assortment optimization under the mixture of multinomial logit models. Oper Res Lett 43(2):189–194. https://doi.org/10.1016/j.orl.2015.01.010

Kunnumkal S, Martínez-de-Albéniz V (2019) Tractable approximations for assortment planning with product costs. Oper Res 67(2):436–452. https://doi.org/10.1287/opre.2018.1771

Lederrey G, Lurkin V, Hillel T, Bierlaire M (2021) Estimation of discrete choice models with hybrid stochastic adaptive batch size algorithms. J Choice Model. https://doi.org/10.1016/j.jocm.2020.100226

Leitner M, Lodi A, Roberti R, Sole C (2023) An exact method for (constrained) assortment optimization problems with product costs. INFORMS J Comput Articles Adv. https://doi.org/10.1287/ijoc.2022.0262

Li X, Ke J (2019) Robust assortment optimization using worst-case CVaR under the multinomial logit model. Oper Res Lett 47(5):452–457. https://doi.org/10.1016/j.orl.2019.07.010

Li G, Rusmevichientong P, Topaloglu H (2015) The d-level nested logit model: assortment and price optimization problems. Oper Res 63(2):325–342. https://doi.org/10.1287/opre.2015.1355

Liu N, Ma Y, Topaloglu H (2020) Assortment optimization under the multinomial logit model with sequential offerings. INFORMS J Comput 32(3):835–853. https://doi.org/10.1287/ijoc.2019.0910

Lo V, Topaloglu H (2021) Omnichannel assortment optimization under the multinomial logit model with a features tree. Manuf Serv Oper Manag 24(2):1220–1240. https://doi.org/10.1287/msom.2021.1001

Luce R (2012) Individual choice behavior: a theoretical analysis. Dover books on mathematics. Dover Publications, New York

Mahajan S, van Ryzin G (2001) Stocking retail assortments under dynamic consumer substitution. Oper Res 49(3):334–351. https://doi.org/10.1287/opre.49.3.334.11210

Maragheh R et al (2021) Choice modeling and assortment optimization in the presence of context effects. SSRN. https://doi.org/10.2139/ssrn.3747354

Martínez-de Albéniz V, Kunnumkal S (2022) A model for integrated inventory and assortment planning. Manag Sci 68(7):5049–5067. https://doi.org/10.1287/mnsc.2021.4149

McElreath MH, Mayorga ME (2012) A dynamic programming approach to solving the assortment planning problem with multiple quality levels. Comput Oper Res 39(7):1521–1529. https://doi.org/10.1016/j.cor.2011.08.023

McElreath MH, Mayorga ME, Kurz ME (2010) Metaheuristics for assortment problems with multiple quality levels. Comput Oper Res 37(10):1797–1804. https://doi.org/10.1016/j.cor.2010.01.011

McFadden D (1978) Modelling the choice of residential location. Transp Res Rec *673*

McFadden D (1973) Conditional logit analysis of qualitative choice behaviour. In: Zarembka P (ed) Frontiers in econometrics. Academic Press, New York, pp 105–142

McFadden D, Train K (2000) Mixed MNL models for discrete response. J Appl Economet 15(5):447–470. https://doi.org/10.1002/1099-1255(200009/10)15:5<447::AID-JAE570>3.0.CO;2-1

Mehrani S, Sefair JA (2022) Robust assortment optimization under sequential product unavailability. Eur J Oper Res 303(3):1027–1043. https://doi.org/10.1016/j.ejor.2022.03.033

Melo M, Nickel S, Saldanha-da Gama F (2009) Facility location and supply chain management—a review. Eur J Oper Res 196(2):401–412. https://doi.org/10.1016/j.ejor.2008.05.007

Méndez-Díaz I, Miranda-Bront JJ, Vulcano G, Zabala P (2014) A branch-and-cut algorithm for the latent-class logit assortment problem. Discrete Appl Math 164(1):246–263. https://doi.org/10.1016/j.dam.2012.03.003

Miao S, Chao X (2020) Dynamic joint assortment and pricing optimization with demand learning. Manuf Serv Oper Manag 23(2):525–545. https://doi.org/10.1287/msom.2019.0857

Miao S, Chao X (2022) Online personalized assortment optimization with high-dimensional customer contextual data. Manuf Serv Oper Manag 24(5):2741–2760. https://doi.org/10.1287/msom.2022.1128

Miao S, Wang Y, Zhang J (2021) A general framework for resource constrained revenue management with demand learning and large action space. SSRN. https://doi.org/10.2139/ssrn.3841273

Miller CM, Smith SA, McIntyre SH, Achabal DD (2010) Optimizing and evaluating retail assortments for infrequently purchased products. J Retail 86(2):159–171. https://doi.org/10.1016/j.jretai.2010.02.004

Mišić VV (2020) Optimization of tree ensembles. Oper Res 68(5):1605–1624. https://doi.org/10.1287/opre.2019.1928

Mišić V, Perakis G (2019) Data analytics in operations management: a review. Manuf Serv Ope Manag 22(1):158–169. https://doi.org/10.1287/msom.2019.0805

Müller S, Haase K (2016) On the product portfolio planning problem with customer-engineering interaction. Oper Res Lett 44(3):390–393. https://doi.org/10.1016/j.orl.2016.03.013

Mushtaque U, Pazour J (2022) Assortment optimization under cardinality effects and novelty for unequal profit margin items. J Revenue Pricing Manag 21:106–126. https://doi.org/10.1057/s41272-020-00279-7

Nip K, Wang Z, Wang Z (2021) Assortment optimization under a single transition choice model. Prod Oper Manag 30(7):2122–2142. https://doi.org/10.1111/poms.13358

Paul A, Feldman J, Davis J (2018) Assortment optimization and pricing under a nonparametric tree choice model. Manuf Serv Oper Manag 20(3):550–565. https://doi.org/10.1287/msom.2017.0662

Peeters Y, den Boer AV (2022) Stochastic approximation for uncapacitated assortment optimization under the multinomial logit model. Nav Res Logist 69(7):927–938. https://doi.org/10.1002/nav.22068

Peeters Y, den Boer AV, Mandjes M (2022) Continuous assortment optimization with logit choice probabilities and incomplete information. Oper Res 70(3):1613–1628. https://doi.org/10.1287/opre.2021.2235

Peng Z, Rong Y, Zhu T (2022) When to sacrifice prediction accuracy: machine learning or MNL choice model for assortment planning. SSRN. https://doi.org/10.2139/ssrn.4298996

Perakis G, Thayaparan L (2023) UMOTEM: upper bounding method for optimizing over tree ensemble models. SSRN. https://ssrn.com/abstract=3972341

Qi M, Mak H-Y, Shen Z-JM (2020) Data-driven research in retail operations—a review. Nav Res Logist 67(8):595–616. https://doi.org/10.1002/nav.21949

Qiu J, Li X, Duan Y, Chen M, Tian P (2020) Dynamic assortment in the presence of brand heterogeneity. J Retail Consum Serv 56:102–152. https://doi.org/10.1016/j.jretconser.2020.102152

Rios I, Saban D, Zheng F (2022) Improving match rates in dating markets through assortment optimization. Manuf Serv Oper Manag 25(4):1304–1323. https://doi.org/10.1287/msom.2022.1107

Rusmevichientong P, Topaloglu H (2012) Robust assortment optimization in revenue management under the multinomial logit choice model. Oper Res 60(4):865–882. https://doi.org/10.1287/opre.1120.1063

Rusmevichientong P, Shen Z-JM, Shmoys DB (2010) Dynamic assortment optimization with a multinomial logit choice model and capacity constraint. Oper Res 58(6):1666–1680. https://doi.org/10.1287/opre.1100.0866

Rusmevichientong P, Shmoys D, Tong C, Topaloglu H (2014) Assortment optimization under the multinomial logit model with random choice parameters. Prod Oper Manag 23(11):2023–2039. https://doi.org/10.1111/poms.12191

Rusmevichientong P, Sumida M, Topaloglu H (2020) Dynamic assortment optimization for reusable products with random usage durations. Manage Sci 66(7):2820–2844. https://doi.org/10.1287/mnsc.2019.3346

Schön C (2010a) On the product line selection problem under attraction choice models of consumer behavior. Eur J Oper Res 206(1):260–264. https://doi.org/10.1016/j.ejor.2010.01.012

Schön C (2010b) On the optimal product line selection problem with price discrimination. Manage Sci 56(5):896–902. https://doi.org/10.1287/mnsc.1100.1160

Schuurman P, Woeginger GJ (2009) Approximation schemes – a tutorial. In: Möhring R, Potts C, Schulz A, Woeginger G, Wolsey L (eds) Lectures on Scheduling

Schwamberger J, Fleischmann M, Strauss A (2023) Tractable time slot assortment optimization in attended home delivery under consider-then-choose customer choice. SSRN. https://doi.org/10.2139/ssrn.4351741

Şen A, Atamtürk A, Kaminsky P (2018) Technical note—a conic integer optimization approach to the constrained assortment problem under the mixed multinomial logit model. Oper Res 66(4):994–1003. https://doi.org/10.1287/opre.2017.1703

Sifringer B, Lurkin V, Alahi A (2020) Enhancing discrete choice models with representation learning. Transportation Research Part B: Methodological 140:236–261. https://doi.org/10.1016/j.trb.2020.08.006

Şimşek S, Topaloglu H (2018) Technical note—an expectation-maximization algorithm to estimate the parameters of the Markov chain choice model. Oper Res 66(3):748–760. https://doi.org/10.1287/opre.2017.1692

Strauss AK, Klein R, Steinhardt C (2018) A review of choice-based revenue management: theory and methods. Eur J Oper Res 271(2):375–387. https://doi.org/10.1016/j.ejor.2018.01.011

Sumida M, Gallego G, Rusmevichientong P, Huseyin T, Davis J (2020) Revenue-utility tradeoff in assortment optimization under the multinomial logit model with totally unimodular constraints. Manage Sci 67(5):2845–2869. https://doi.org/10.1287/mnsc.2020.3657

Talluri K, van Ryzin G (2004) Revenue management under a general discrete choice model of consumer behavior. Manage Sci 50(1):15–33. https://doi.org/10.1287/mnsc.1030.0147

Thomé AMT, Scavarda LF, Scavarda AJ (2016) Conducting systematic literature review in operations management. Prod Plan Control 27(5):408–420. https://doi.org/10.1080/09537287.2015.1129464

Torrico A, Carvalho M, Lodi A (2021) Multi-agent assortment optimization in sequential matching markets. arXiv. https://doi.org/10.48550/arXiv.2006.04313

Train KE (2009) Discrete choice methods with simulation, 2nd edn. Cambridge University Press, Cambridge

Transchel S, Buisman M, Haijema R (2022) Joint assortment and inventory optimization for vertically differentiated products under consumer-driven substitution. Eur J Oper Res 301(1):163–179. https://doi.org/10.1016/j.ejor.2021.09.041

Tulabandhula T, Sinha D, Karra S (2022) Optimizing revenue while showing relevant assortments at scale. Eur J Oper Res 300(2):561–570. https://doi.org/10.1016/j.ejor.2021.08.006

Tulabandhula T, Sinha D, Karra S, Patidar P (2023) Multi-purchase behavior: modeling, estimation, and optimization. Manuf Serv Oper Manag 25(6):2298–2313. https://doi.org/10.1287/msom.2020.0238

Udwani R (2021) Submodular order functions and assortment optimization. arXiv. https://doi.org/10.48550/arXiv.2107.02743

van Cranenburgh S, Wang S, Vij A, Pereira F, Walker J (2022) Choice modelling in the age of machine learning—discussion paper. J Choice Model 42:1003. https://doi.org/10.1016/j.jocm.2021.100340

Wang R (2018) When prospect theory meets consumer choice models: assortment and pricing management with reference prices. Manuf Serv Oper Manag 20(3):583–600. https://doi.org/10.1287/msom.2017.0688

Wang R (2021) Technical note—consumer choice and market expansion: modeling, optimization, and estimation. Oper Res 69(4):1044–1056. https://doi.org/10.1287/opre.2020.2059

Wang R, Sahin O (2017) The impact of consumer search cost on assortment planning and pricing. Manag Sci 64(8):3649–3666. https://doi.org/10.1287/mnsc.2017.2790

Wang R, Wang Z (2016) Consumer choice models with endogenous network effects. Manag Sci 63(11):3944–3960. https://doi.org/10.1287/mnsc.2016.2520

Wang S, Mo B, Hess S, Zhao J (2021) Comparing hundreds of machine learning classifiers and discrete choice models in predicting travel behavior: an empirical benchmark. arXiv. https://doi.org/10.48550/arXiv.2102.01130

Wang C, Wang Y, Tang S (2022a) When advertising meets assortment planning: joint advertising and assortment optimization under multinomial logit model. SSRN. https://doi.org/10.2139/ssrn.3908616

Wang R, Zhao Z, Ke C (2022b) Modeling consumer choice and optimizing assortment under the threshold multinomial logit model. SSRN. https://doi.org/10.2139/ssrn.4184044

Wang H, Cai Z, Li X, Talluri K (2023a) A neural network based choice model for assortment optimization. arXiv. https://doi.org/10.48550/arXiv.2308.05617

Wang H, Li X, Talluri K (2023b) Transformer choice net: a transformer neural network for choice prediction. arXiv. https://doi.org/10.48550/arXiv.2310.08716

Wang X, Wei M, Yao T (2023c) Online assortment optimization with high-dimensional data. SSRN. https://doi.org/10.2139/ssrn.3521843

Wang M, Zhang X, Li X (2023d) Multiple-purchase choice model: estimation and optimization. Int J Prod Econ 265. https://doi.org/10.1016/j.ijpe.2023.109010

Wang M, Zhang H, Rusmevichientong P, Shen Z-JM (2023e) Optimizing offline product design and online assortment policy: measuring the relative impact of each decision. SSRN. https://doi.org/10.2139/ssrn.4090147

Whitley D (2013) Sharpened and Focused No Free Lunch and Complexity Theory, Ch. 16 in Search Methodologies: Introductory Tutorials in Optimization and Decision Support Techniques, 451–476 Springer New York, NY

Xie T, Ge D (2018) A tractable discrete fractional programming: application to constrained assortment optimization. J Comb Optim 36:400–415. https://doi.org/10.1007/s10878-018-0302-x

Zhang H, Rusmevichientong P, Topaloglu H (2020) Assortment optimization under the paired combinatorial logit model. Oper Res 68(3):741–761. https://doi.org/10.1287/opre.2019.1930