European Journal of Operational Research 285 (2020) 1058–1076

Contents lists available at ScienceDirect

European Journal of Operational Research

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

Innovative Applications of O.R.

Product assortment and space allocation strategies to attract loyal and non-loyal customers

Anna Timonina-Farkas[a,*], Argyro Katsifou[a], Ralf W. Seifert[a,b]

[a] College of Management of Technology (CDM), École Polytechnique Fédérale de Lausanne (EPFL), EPFL CDM TOM, ODY (Odyssea), Station 5, Lausanne 1015, Switzerland
[b] International Institute for Management Development (IMD), Chemin de Bellerive 23, Lausanne 1003, Switzerland

ARTICLE INFO

Article history:
Received 17 December 2018
Accepted 10 February 2020
Available online 19 February 2020

Keywords:
OR in marketing
Assortment planning
Inventory management
Bilevel optimization
Stochastic optimization

ABSTRACT

Assortment planning deserves much attention from practitioners and academics due to its direct impact on retailers' commercial success. In this paper we focus on the increasingly popular retail practice to use combined product assortments with both "standard" and more fashionable and short-lived "variable" products for building up store traffic of "loyal" and "non-loyal" heterogeneous customers and enlarging the sales due to the potential cross-selling effect.

Addressing the assortment planning as a bilevel optimization problem, we focus on decision-dependent uncertainties: the retailer's binary decision about product inclusion influences the distribution of the product's demand. Furthermore, our model accounts for customers' optimal purchase quantities, which depend on budget constraints limiting the basket that a customer is able to purchase.

We propose iterative heuristics using optimal quantization of demand and customers budget distributions to define the total assortment and the inventory level per product. These heuristics provide lower bounds on the optimal value. We conduct a comparison to other existing lower bounds and we formulate upper bounds via linear (LP) and semidefinite (SDP) relaxations for the performance evaluation of the heuristics and for an efficient numerical solution in high-dimensional cases.

For managerial insights, we compare the proposed approach with three assortment planning strategies: (1) the retailer does not carry variable products; (2) the retailer ignores the cross-selling effect; and (3) the maximum space allocated to each product is fixed. Our results suggest that variable assortment boosts the retailers profits if the cross-selling effect is not neglected in the decision about products quantities.

1. Introduction

Product assortment decisions have a direct impact on a retailer's commercial success. The goal of assortment optimization is to specify which products and in which quantities should be included in the assortment to maximize the retailer's profit subject to various constraints, such as limited shelf space for displaying products or budget restrictions.

Research (Davenport & Harris, 2007; Liberatore & Luo, 2010; Ranyard, Fildes, & Hu, 2015; SAS Institute, 2013) indicates that customer analytics focusing on customer economics, personalized offers and customer-centric merchandising identify profitable growth opportunities for retailers. These opportunities go undetected otherwise. In this paper, we propose a stylized OR model for decision support which combines customer-centric merchandising with customer economics insights.

First, maintaining customer loyalty for customer-centric merchandising, we focus on the retailer's assortment strategies driven by the increasingly important commercial distinctions between "standard" and "variable" products serving "loyal" and "non-loyal" customers: standard assortment is a strategic tool to attract and retain loyal customers who regularly visit the store for their customary shopping and grow accustomed to a certain product range (Billington & Nie, 2009; Grewal, Levy, Mehrotra, & Sharma, 1999; Walters & Hanrahan, 2000); variable products are critical for the increase of the overall store traffic by attracting non-loyal customers who are rather opportunistic and are occasionally attracted by special offers (Katsifou, 2013; Katsifou, Seifert, & Tancrez, 2014). Next, implementing customer economics insights to the model, we work with a common lifestyle and life-stage customer segment "Shoppers on a Budget" (SAS Institute, 2013) by implementing customers' budget constraints to the model. Each customer's budget is

[*] Corresponding author.
E-mail addresses: anna.farkas@epfl.ch (A. Timonina-Farkas), argyro.katsifou@gmail.com (A. Katsifou), ralf.seifert@epfl.ch (R.W. Seifert).

https://doi.org/10.1016/j.ejor.2020.02.019

Table 1
Variable assortment characteristics.

Retailer	Introduction	Sales Period			% of the total
	Frequency	Week	Month	> Month	assortment
Walmart	Weekly	√			1
Target	1.5-3 months			√	N/A
Lidl	Twice/week	√			9
Aldi	Twice/week	√			4
Zara	Twice/week		√		75

uncertain to the retailer, who can only estimate the budget distribution. To the best of our knowledge, current literature focusing on the retailer's product assortment considers the vendor's shelf space and his budget constraints while constantly neglecting customers' budget limitations due to their modeling complexity. In this work, we create a bilevel model, where the retailer's optimal product assortment decision directly depends on the customers' optimal purchase decision, which, in turn, depends on the customers' budget distribution.

Many retailers strive to take advantage of a combined assortment of standard and variable products independently of a retailer's product variety strategy. The assortment of standard products is relatively stable over time, is not subject to seasonality and is expected to be present at every time customers shop (Billington & Nie, 2009). The main objective of this assortment is to support the positioning and profitability of a retailer (Katsifou et al., 2014; Walters & Hanrahan, 2000). Variable products, by contrast, are seasonal, they are offered only temporarily, in limited quantities and usually at very attractive prices. Variable products may provide retailers with profits or even be loss-making if considered in isolation (Katsifou, 2013; Katsifou et al., 2014), but they are important because they help to increase the number of customers in a store: this increased store traffic can influence sales of standard products, resulting in expansion of overall sales. Attracting customers drawn to promotions, retailers provide them with a sense of achievement by proposing items on special offer (Marquard, 2007). Doing this by rotating variable assortment items might be preferable to price promotions of standard assortment items, since it allows retailers to keep a steady set of standard products and helps to maintain a consistent pricing over the long duration avoiding erosion of established price points. Moreover, the cross-selling effect, which occurs for both groups of customers, encourages loyal customers to purchase variable products once in the store, and makes non-loyal customers purchase standard products, increasing total profits.

As an example, Walmart offers a wide range of standard product categories - from food and clothing to home electronics - as well as a limited number (around 1% of the total assortment size) of weekly "special buys" in order to create excitement or to clear out stock (Marquard, 2007). "Limited-time-only deals" are also a key part of US-based retailer Target's strategy to attract shoppers. These special items are usually offered for six-week to 90-day periods. By contrast, Lidl and Aldi, two of the most successful discount retailers, compete effectively with a very narrow product assortment. Their assortment consists of limited standard products, but they continuously present variable assortment articles (called "surprise buys" or "special items"), consisting of very diverse non-food products (e.g., "back to school" special items or ski wear), with the aim of keeping the shopping experience in the store special (Kumar & Steenkamp, 2007; Zentes, Morschett, & Schramm-Klein, 2007). Table 1 summarizes the aforementioned assortment strategies.

The rest of the paper is organized as follows. Section 2 provides a literature review on the topic. In Section 3, we present the problem and the mathematical model. In Section 4, we formulate lower and upper bounds on our bilevel optimization problem, proposing heuristic solution procedures taking the dependency between product demands and the total assortment decision into account. In Section 5, we present and discuss illustrative examples and insights: we compare the proposed approach with three different assortment planning strategies: (1) in the first one, the retailer does not carry variable products; (2) in the second, the retailer ignores the cross-selling effect; and (3) in the third, the space allocated to each type of product is fixed a priori.

2. Literature review

Works (Kök, Fisher, & Vaidyanathan, 2009; Mou, Robb, & DeHoratius, 2018) provide the product assortment literature review: studies that deal with the total assortment problem, examining the strategic role of key product categories, are quite rare (Amit, Mehta, & Tripathi, 2015; Ghoniem & Maddah, 2015; Katsifou, 2013; Katsifou et al., 2014; Rabbani, Salehi, & Farshbaf-Geranmayeh, 2017). In our research, we devise a model to optimize the total combined assortment of a retailer, taking into consideration the role of both standard and variable assortments in generating store traffic. The problem we study requires to explicitly model the interdependence between the demand for a particular product and the total assortment decision: the demand depends on characteristics of customers attracted to the store, such as their attitude towards risk and their budget; in turn, customers' optimal purchase quantities depend on the selected assortment (standard or variable). Taking this interdependence into account in product assortment decisions is one of the main contributions of our work. Indeed, to the best of our knowledge, prior studies did not investigate the influence of an uncertain customers' budget and customers' optimal purchase quantities on the store's optimal assortment. The work (Katsifou et al., 2014), where the authors construct a stylized model for customers' behavior based on their purchase baskets, is in the same research stream: however, in Katsifou et al. (2014), each customer is assumed to buy only one item of a product and neither the customers' optimal purchase quantities nor their budget constraints are taken into account. In our work, we construct and solve the bilevel optimization problem, explicitly accounting for the interdependency between optimal demand and assortment and addressing the asymmetrical influence of different products on a store's profits. Our solution approach is novel to the existing literature, it allows to introduce both lower and upper bounds on the optimal profit and to handle large-scale problems.

Apart from looking at the retailers' total product assortment selection, we also explicitly tackle the inventory management of selected products with a limited shelf space (Hübner, Kuhn, & Kühn, 2016). Product assortment and inventory management are two naturally interconnected and critically important retailing decisions that have a substantial impact on retailers' profitability (Kök & Fisher, 2007). Yet product assortment planning and multi-item inventory management are two literature streams that have developed somewhat independently. Article (Kök et al., 2009) provides a detailed review of the product assortment planning literature. One of the first papers in the operations research literature integrating inventory management into the product assortment planning problem is van Ryzin (1999). The article studies the trade-off between product variety benefits and inventory costs using a Multinomial Logit (MNL) model.

The MNL model is a widely used utility-based model where each consumer chooses the product with the highest utility out of the set of available choices (Ben-Akiva & Lerman, 1994). In van Ryzin (1999), customers have homogeneous expected utilities and can substitute if their favorite variant is not carried (assortment-based substitution). The sale is lost if their favorite variant is carried but temporarily unavailable (no stock-out-based substitution). Through the use of the newsvendor framework, they compute the inventory level per product as well as the optimal assortment, which is shown to consist of a certain number of the most popular products. In the same research stream, article (Li, 2007) studies a joint product assortment and inventory optimization problem for a single period, providing a closed-form solution to determine the optimal assortment when the store traffic is treated as a continuous random variable, and a heuristic for discrete store traffic. Both studies assume that consumers are homogeneous with respect to their preferences. This assumption is relaxed in Katsifou et al. (2014), Katsifou (2013) and Mahajan and van Ryzin (2001). The work Katsifou et al. (2014) is focused on loyal/non-loyal customers and provides an assortment optimization model describing consumer preferences and the choice of retailer in the MNL framework (Anderson, de Palma, & Thisse, 1992).

Whereas (Mahajan & van Ryzin, 2001) studies individual heterogeneity, we, similar to Katsifou et al. (2014), are focused on "customer type" heterogeneity. However, in contrast to a so-called latent-class logit assortment problem, where a consumer belongs to a market segment characterized by a "consideration set", i.e., the set of products that the customer is considering purchasing, we do not bound product assortments available for loyal and non-loyal customers. To model different attitudes towards standard and variable products, we assume different risk-bearing abilities towards them for loyal and non-loyal customers. Before, the work (Méndez-Diaz, Bront, Vulcano, & Zepala, 2010) studied the latent-class logit assortment problem, where segments were allowed to overlap, i.e., a product could belong simultaneously to the consideration sets of two or more segments. They proposed a branch-and-bound algorithm to find the optimal solution to the constrained and the unconstrained versions of the problem. In our research, we seek to optimize the total assortment considering a joint retail space constraint (Hübner et al., 2016), whereas (Méndez-Diaz et al., 2010) examined only a single product category optimization problem, and their model was constrained by the number of products to be displayed.

In another stream of research on product assortment planning, exogenous demand models are used to model consumer choice. These models specify a priori the demand for each product and the probability that an individual will choose another product as a substitute when his favorite product is not available. Compared with the MNL model used in Katsifou et al. (2014), the exogenous demand model is more flexible in dealing with both assortment-based and stock-out-based substitutions but requires more effort in data collection and parameter estimation. Article (Smith & Agrawal, 2000) deploys an exogenous demand model, by assuming that demand follows a negative binomial distribution (Agrawal & Smith, 1996); the authors estimate substitution probabilities and overall cross-buy schemes of different products to populate a matrix. Apart from assortment optimization, they use an approximate newsvendor model to compute the optimal stock level for each product, subject to an assortment fill rate constraint. Article (Agrawal & Smith, 2003) extends this work and incorporates the complementarity and the substitution effects in a set of products demanded by a customer. In our research, we do not consider substitutions but we make a step forward in modeling optimal demand: we formulate and solve a bilevel optimization problem (Bard, 1991; 1998; Ben-Ayed & Blair, 1990; Calvete, Galé, & Mateo, 2008; Colson, Marcotte, & Savard, 2007), where the retailer's first-level binary decision about product inclusion influences the distribution of the product's demand driven by the second-level customer choice model. Each customer has his own optimization problem to solve and the optimal solution depends on the assortment available at the retailer. The optimal solution of the customer's optimization problem is the quantity to purchase under the budget constraint. We model the customer's choice of retailer via a simplistic model (Anderson et al., 1992; Ben-Akiva & Lerman, 1994; Smith & Agrawal, 2000) based on market shares. This allows to avoid the MNL model independency assumption and the assumption about total number of retailers.

Overall, we consider "loyal" and "non-loyal" customers in our bilevel optimization. The cross-selling effect may occur for both groups of customers, encouraging loyal customers to purchase variable products, as well as making non-loyal customers purchase standard products. We describe the strength of the cross-selling effect via coefficients of relative risk aversion towards standard and variable products for different groups of customers. Furthermore, we distinguish between budget distributions and budget constraints for each group of customers, limiting baskets that different customers are able to purchase.

We propose iterative heuristics to define the total assortment and inventory level per product, taking the dependency between the demand for a particular product and the total assortment decision explicitly into account. These heuristics provide lower bounds on the optimal value of the optimization problem. We also formulate two upper bounds for the bilevel optimization problem via linear (LP) and semidefinite (SDP) relaxations (Fortet, 1960; Gaivoronski, Lisser, & Lopez, 2011). These bounds are useful for the performance evaluation of the heuristics, as well as for the efficiency of solution estimation.

We use optimal quantization techniques (Monge, 1781; Rachev & Rüschendorf, 1998; Villani, 2008) instead of typical Monte-Carlo simulations for the approximation of demand and customer's budget distributions. This allows to enhance both accuracy and efficiency of the numerical solution, as the number of quantizers for an accurate estimate can be much lower than in Monte-Carlo sampling (Timonina, 2013). We consider optimality of scenario quantization methods in the sense of minimal Kantorovich-Wasserstein distance (Kantorovich, 1942), which allows to implement the structural information in order to take more accurate decisions, as well as to bound the approximation error.

3. Problem and model formulation

We consider a retailer seeking to optimize the product assortment and the inventory level per product in a store. We suppose that retailer r selects the assortment from a set of candidate products. Product substitution is not considered and stock-outs result in lost sales (Katsifou, 2013; Katsifou et al., 2014). With this product assortment decision, the goal of the retailer is to attract as large as possible share of the customer base, i.e., as many potential customers as possible in a certain geographical area. In particular, the two product groups are primarily targeted at different customer segments. In our model we separate the total customer base into two segments: *loyal* and *non-loyal* customers represented by the sets $\mathcal{L}$ and $\overline{\mathcal{L}}$, respectively. Loyal customers are attracted to the store primarily by standard products (e.g., Aldi's loyal customers do their everyday shopping there mainly because standard products satisfy their needs). On the other hand, non-loyal customers look for special offers and are attracted to the store primarily by variable assortments (e.g., Aldi's non-loyal customers read the weekly flyer with the "surprise buys" and decide whether to visit the store or not).

Though loyal and non-loyal customers are attracted to the store by standard and variable products respectively, they may decide, once in the store, to purchase products from both product groups. For instance, Aldi's loyal customer who arrives at the store to buy his customary standard products, may also buy some variable products like a garden chair at an attractive price. On the other hand, non-loyal customers that visit the store to buy variable products, may also decide, once in the store, to take the opportunity to cover their daily needs by buying some standard products like milk and orange juice. The fact that standard (respectively variable) products are bought by non-loyal (respectively loyal) customers is commonly known as cross-selling effect.

The retailer has to resolve the trade-off between store attractiveness and the corresponding operational costs. By increasing the number of products in the assortment, a higher number of customers can be attracted to the store, driving up traffic and consequently the retailers' overall sales. On the other hand, the introduction of a product to the assortment comes with a setup order cost as well as related inventory costs. Furthermore, as the shelf space is limited, a larger number of products also means less storage space and consequently lower inventory level per product while the demand increases, leading to higher inventory costs. The low inventory level may result in an increase in the stock-out probability or in frequent replenishment (for standard products) and thus a higher setup cost. The retailer has to find a trade-off between standard and variable products, taking into consideration the shelf space constraint. In general, the main objective of stocking variable products is to attract non-loyal customers who will then buy both variable and standard products.

The inventory policies vary between the two groups of products. Standard products are carried over a long time horizon, and the retailer has to pay a holding cost for the units stored at the end of each sales period. Variable products, by contrast, are rotated frequently, resulting in a sequential one-period problem. At the end of their sales period the unsold variable products cause an overstock cost. Our mathematical model described further investigates the aforementioned trade-offs.

The remainder of this section is structured as follows: first, we develop the retailer's profit functions per product group with respect to its own characteristics such as the sales period and the inventory costs; second, we explicitly model the demands per product taking into consideration customers' store choice and their optimal in-store purchase decision; finally, using the demand as an input to the profit function, we give the mathematical formulation to maximize the retailer's total profit subject to the shelf space constraint. We formulate a bilevel optimization problem, whose aim is not only to maximize the retailer's total profit but also to investigate optimal in-store purchase decisions of customers.

3.1. Profit functions

We assume that standard products are carried by the retailer over a long time horizon. A periodic review order-up-to inventory policy (Nahmias, 2005) is used. The inventory of a standard product j is replenished every period T^s_j to reach the order-up-to inventory level Q^s_j. The lead time is assumed to be negligible. The period T^s_j is defined prior to the optimization process for each product j while the inventory level Q^s_j is a decision variable in our optimization model.

To construct the profit function for standard products, one needs to account for uncertain demand for these products. For this, let $d^s_j = (d^s_{j1}, d^s_{j2}, \dots, d^s_{jt}, \dots, d^s_{jT^s_j}, \dots)$ be the stochastic demand process described by probability distribution functions $F^s_{jt}(x) = \mathbb{P}(d^s_{jt} \le x)$ and corresponding density functions $f^s_{jt}(x)$. For standard products, one can reasonably assume that random demands $(d^s_{j1}, d^s_{j2}, \dots, d^s_{jt}, \dots, d^s_{jT^s_j}, \dots)$ are independent normally distributed random variables with mean μ^s_j and standard deviation $\sqrt{\mu^s_j}$, approximating the Poisson distribution (van Ryzin, 1999), i.e., $F^s_{jt}(x) = F^s_j(x),\ \forall t$. We denote by $D^s_j = \sum_{t=1}^{T^s_j} d^s_{jt}$ the cumulative demand for the standard product j during period T^s_j with probability distribution $G^s_j(y) = \mathbb{P}(D^s_j \le y)$ and corresponding density function $g^s_j(y)$. As we assume random variables $(d^s_{j1}, \dots, d^s_{jt}, \dots, d^s_{jT^s_j}, \dots)$ to be independent and normally distributed, we also know the distribution of the variable D^s_j: the probability distribution $G^s_j(y) = \mathbb{P}(D^s_j \le y)$ is normal with mean $T^s_j\mu^s_j$ and standard deviation $\sqrt{T^s_j\mu^s_j}$.

Next, let K^s_j be the setup cost per positive order planned and let c^s_j be the proportional order cost per unit ordered. In this case, the total fixed plus proportional order cost is $K^s_j + c^s_j\min\{D^s_j,\ Q^s_j\}$, where Q^s_j is the order-up-to level. As the inventory level at the end of the cycle is $\max\{Q^s_j - D^s_j,\ 0\}$, the holding cost from cycle to cycle is equal to $h^s_j\max\{Q^s_j - D^s_j,\ 0\}$ with h^s_j being the proportional holding cost per unit stored. Further, if the demand exceeds the level of inventory, the opportunity costs need to be taken into account. These costs are equal to $u^s_j\max\{D^s_j - Q^s_j,\ 0\}$ with u^s_j denoting the loss-of-goodwill cost per unit. Also, let p^s_j be the selling price per unit.

We denote the profit margin by $m^s_j = p^s_j - c^s_j$, set $cu^s_j = m^s_j + h^s_j + u^s_j$, which clearly satisfies $cu^s_j > h^s_j$ and $cu^s_j > m^s_j$, and use the total demand mean and variance $\mathbb{E}(D^s_j) = T_j\mu^s_j$ and $\mathbb{V}\mathrm{ar}(D^s_j) = T_j\mu^s_j$ for our Gaussian approximation of the Poisson distribution with $\mu^s_j = \mathbb{E}(d^s_{jt}),\ \forall t$ (van Ryzin, 1999). The expected profit, brought by the standard product $j \in \mathcal{N}_s$ over one time unit, represents a lost-sales model with zero lead time and, according to Appendix 1.1, is equal to

$$\mathbb{E}\left(\frac{\Pi^s_j}{T^s_j}\right) = m^s_j\mu^s_j - \left[\frac{K^s_j}{T^s_j} + \frac{h^s_j}{T^s_j}\left(Q^s_j - T^s_j\mu^s_j\right) + \frac{cu^s_j}{T^s_j}\int_{Q^s_j}^{\infty}(y - Q^s_j)g^s_j(y)dy\right]. \tag{1}$$

Compared to standard products, variable products have a significantly shorter sales period (e.g., one week in Aldi). For this reason, we suppose that the products are ordered at the beginning of the sales period, with no replenishment opportunity. This period is denoted T^v_j and is supposed to be fixed externally, e.g., for marketing reasons. Based on these characteristics, we naturally use a newsvendor-like inventory policy (Nahmias, 2005) to derive the inventory costs of a variable product j.

The inventory costs are similar to those encountered for the standard products. A stock-out cost accounts for the lost sales (Bijvank & Vis, 2011), and a setup cost is paid by the retailer for introducing a new variable product. However, an over-stock cost replaces the holding cost as the variable products are removed from the assortment at the end of a period and cannot be sold in the next one.

Analogous to standard products, let $d^v_j = (d^v_{j1}, d^v_{j2}, \dots, d^v_{jt}, \dots, d^v_{jT^v_j})$ be the stochastic demand process described by probability distribution functions $F^v_{jt}(x) = \mathbb{P}(d^v_{jt} \le x)$ with known parameters $\forall t$ and corresponding density functions $f^v_{jt}(x)$. Note that $t \in \{1, 2, \dots, T^v_j\}$ is finite in this case. As before, we introduce cumulative demand $D^v_j = \sum_{t=1}^{T^v_j} d^v_{jt}$ for the variable product j during period T^v_j with probability distribution $G^v_j(y) = \mathbb{P}(D^v_j \le y)$ and corresponding density function $g^v_j(y)$. Although for standard products one can reasonably assume that random demands $(d^s_{j1}, d^s_{j2}, \dots, d^s_{jt}, \dots, d^s_{jT^s_j}, \dots)$ are independent identically distributed random variables, for variable products a similar assumption for vector $(d^v_{j1}, d^v_{j2}, \dots, d^v_{jT^v_j})$ should generally be relaxed, particularly as the retailer aims to influence the demand by introduction of these products. Furthermore, variable products are likely to be very seasonal or at early stages of their life cycle where demand distribution is changing. To account for this, one would need to employ coupling techniques to estimate the density of the sum of dependent random variables (Hochrainer-Stigler, Timonina-Farkas, Silm, & Balkovič, 2019). If random variables are independent but not necessarily identically distributed, one uses independent convolution to estimate the density $g^v_j(y)$. However, due to the lack of data on variable products, we still use the assumption that random variables $(d^v_{j1}, d^v_{j2}, \dots, d^v_{jt}, \dots, d^v_{jT^v_j})$ are independent normally distributed variables with known mean μ^v_j and standard deviation $\sqrt{\mu^v_j}$, approximating the Poisson distribution (van Ryzin, 1999), i.e., $F^v_{jt}(x) = F^v_j(x),\ \forall t$.

Next, let K^v_j be the setup cost per positive order planned and let c^v_j be the proportional order cost per unit of a variable product ordered. In this case, the total fixed plus proportional order cost is $K^v_j + c^v_j Q^v_j$, where Q^v_j is the order placed. Also, let us denote by co^v_j the over-stock cost per unit. For variable products, there is no holding cost from cycle to cycle and the total value of over-stock $co^v_j \max\{Q^v_j - D^v_j,\ 0\}$ is being lost. The loss-of-goodwill cost is equal to $u^v_j \max\{D^v_j - Q^v_j,\ 0\}$ with u^v_j denoting the loss-of-goodwill cost per unit. Let p^v_j be the selling price per unit. Derived in Appendix 1.1, the expected profit brought by the variable product $j \in \mathcal{N}_v$ over one time unit is equal to

$$\mathbb{E}\left(\frac{\Pi^v_j}{T^v_j}\right) = m^v_j \mu^v_j - \left[\frac{K^v_j}{T^v_j} + \frac{c^v_j}{T^v_j}\left(Q^v_j - T^v_j \mu^v_j\right) + \frac{cu^v_j}{T^v_j}\int_{Q^v_j}^{\infty} (y - Q^v_j) g^v_j(y)dy + \frac{co^v_j}{T^v_j}\int_0^{Q^v_j} (Q^v_j - y) g^v_j(y)dy\right], \tag{2}$$

where we denote the profit margin by $m^v_j = p^v_j - c^v_j$ and set $cu^v_j = m^v_j + c^v_j + u^v_j$ and where $cu^v_j > c^v_j$ and $cu^v_j > m^v_j$.

Note that the profit function of a variable product j (Eq. (2)) is defined on its sales period T^v_j while the profit function of a standard product (Eq. (1)) is defined on a longer time horizon T^s_j. This timing difference can be explained in two ways when considering the assortment decision (and thus when applying our approach as described below). Either the assortment is decided every period T^v_j depending on the potential variable products available for the particular period. Or the assortment is decided for a longer horizon and it is supposed that the potential variable products, as a group, keep the same characteristics (e.g., attractiveness and volume) in each period (and thus, even if different in nature, the selected variable products keep the same characteristics throughout the time horizon). The first option gives a more stringent assortment decision but requires to adapt the assortment very frequently. The second option gives a rather good picture of the strategic level of decision (space sharing, number of products, general characteristics, etc.).

In the rest of our work, we refer to the second assumption by default but our approach can also be used in the first case, to optimize the assortment at every sales period T^v_j.

3.2. Demand model

Profit functions depend on mean demands of standard and variable products, which can be influenced by the retailer's assortment decision. We derive mean demands per time unit based on the retailer's assortment decision, customers' choice of the retailer and in-store product choice. The total set of candidate products is divided into two groups: the standard products, $\mathcal{N}_s = \{1, \dots, N_s\}$, and the variable products, $\mathcal{N}_v = \{1, \dots, N_v\}$. The set of products available at the retailer is known to customers and is represented by the binary vectors $\boldsymbol{s} = \left(S_1, \dots, S_{N_s}\right)$ and $\boldsymbol{v} = \left(V_1, \dots, V_{N_v}\right)$, respectively. The products' quantities are unknown to customers due to customers' inability to observe the backroom.

Customer i solves the following optimization problem to decide about purchase quantities ζ^s_{ij} and ξ^v_{ij} of standard $j \in \mathcal{N}_s$ and variable $j \in \mathcal{N}_v$ products:

$$\begin{aligned} U^*_i = \max_{\zeta^s_{ij}, \xi^v_{ij}} & \sum_{j \in \mathcal{N}_s} u^s_{ij}(\zeta^s_{ij}) + \sum_{j \in \mathcal{N}_v} u^v_{ij}(\xi^v_{ij}), \\ \text{subject to } & \sum_{j \in \mathcal{N}_s} p^s_j \zeta^s_{ij} + \sum_{j \in \mathcal{N}_v} p^v_j \xi^v_{ij} \le B_i, \\ & 0 \le \zeta^s_{ij} \le MS_j,\ \forall j \in \mathcal{N}_s, \\ & 0 \le \xi^v_{ij} \le MV_j,\ \forall j \in \mathcal{N}_v, \end{aligned} \tag{3}$$

where big M is a large enough number, B_i is the customer's budget, $p^s_j,\ p^v_j$ are the products' prices, utilities $u^s_{ij},\ u^v_{ij}$ are the functions describing risk-bearing ability of customer i and U^*_i is the optimal value.

We use isoelastic functions $u^s_{ij}(\zeta^s_{ij}) = \frac{(\zeta^s_{ij})^{1-\gamma_{ij}}}{1-\gamma_{ij}}$, $u^v_{ij}(\xi^v_{ij}) = \frac{(\xi^v_{ij})^{1-\delta_{ij}}}{1-\delta_{ij}}$ with $\gamma_{ij} \in [0,1)$, $\delta_{ij} \in [0,1)$ and exponential utilities $u^s_{ij}(\zeta^s_{ij}) = 1 - \exp(-\gamma_{ij}\zeta^s_{ij})$, $u^v_{ij}(\xi^v_{ij}) = 1 - \exp(-\delta_{ij}\xi^v_{ij})$ with $\gamma_{ij},\ \delta_{ij} > 0\, \forall i, j$ to describe the behavior of risk-averse customers. Using isoelastic utility, one can distinguish between preferences for small and large product quantities, as, increasing the risk-aversion parameter, one observes an increase in the utility towards small quantities and a decrease in the utility towards larger amounts. Differently, incorporating the exponential utility for modeling risk-averse customers, one can clearly rank customers by their preferences: customers with a higher risk-bearing ability parameter γ_{ij} have higher utilities, converging to one in the limit. Further, the risk-neutral set of customers has linear utility functions and the risk-loving group of customers can be described by the utilities $u^s_{ij}(\zeta^s_{ij}) = \frac{1-\exp(-\gamma_{ij}\zeta^s_{ij})}{\gamma_{ij}}$, $u^v_{ij}(\xi^v_{ij}) = \frac{1-\exp(-\delta_{ij}\xi^v_{ij})}{\delta_{ij}}$ with $\gamma_{ij},\ \delta_{ij} < 0$. Modeling risk-neutral and risk-loving preferences of customers would result in no diversification between products: the customer selects a product with the lowest price.

Given the utility function, parameters γ_{ij} and δ_{ij} describe customer i's attitude towards the product j for standard and variable products correspondingly. In this article, we distinguish between following coefficients of relative risk aversion (or, risk-bearing abilities in case of exponential utilities), i.e.,

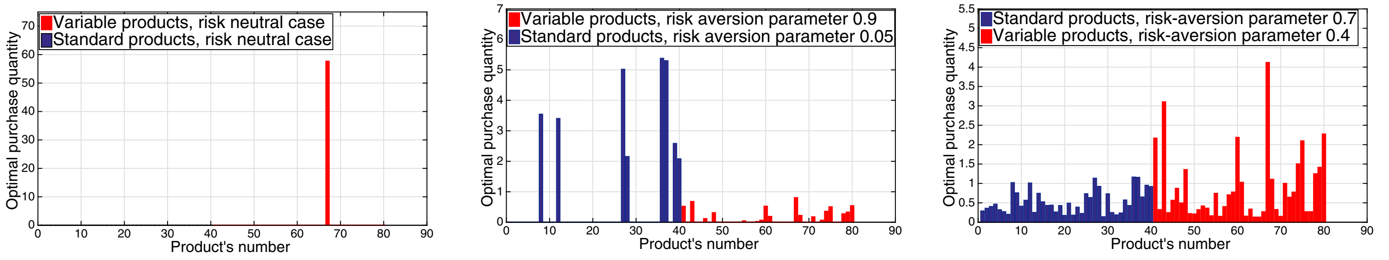
$$\gamma_{ij} = \begin{cases} \gamma, & \forall i \in \mathcal{L},\ j \in \mathcal{N}_s \\ \overline{\gamma}, & \forall i \in \overline{\mathcal{L}},\ j \in \mathcal{N}_s \end{cases} \quad \text{and} \quad \delta_{ij} = \begin{cases} \delta, & \forall i \in \mathcal{L},\ j \in \mathcal{N}_v \\ \overline{\delta}, & \forall i \in \overline{\mathcal{L}},\ j \in \mathcal{N}_v. \end{cases} \tag{4}$$

One could easily generalize the approach and distinguish between coefficients of relative risk aversion for other customer segments and product groups.

Further, the budget B_i is known to the customer i, but unknown to the retailer, who supposes B_i to be distributed in line with the lognormal distribution function $H_i(x) = \mathbb{P}(B_i \le x)$ with parameters μ_i and σ^2_i (Aitchison & Brown, 1957; Prais & Houthakker, 1971). We account for budget distributions that may vary for different groups of customers, i.e.,

$$\mu_i = \begin{cases} \mu, & \forall i \in \mathcal{L} \\ \overline{\mu}, & \forall i \in \overline{\mathcal{L}} \end{cases} \quad \text{and} \quad \sigma^2_i = \begin{cases} \sigma^2, & \forall i \in \mathcal{L} \\ \overline{\sigma}^2, & \forall i \in \overline{\mathcal{L}}. \end{cases} \tag{5}$$

By this, the retailer can better investigate how to develop strategies to attract targeted types of customers from different budget groups. Using problem (3), one can account for different customer segments by varying risk aversion and budget parameters. We model loyal customers $i \in \mathcal{L}$ who are regular at the retailer by



a) Risk-neutral customers: no diversification.

b) Risk-averse customer preferring standard products.

c) Risk-averse customer preferring variable products.

Fig. 1. Optimal purchase quantities for different types of customers.

a lower risk aversion (or, a higher risk-bearing ability in case of exponential utilities) for standard products than for variable ones. Further, we model non-loyal customers $i \in \overline{\mathcal{L}}$ who are in general opportunistic by a higher risk aversion for standard products than for variable ones. Also, as specified before, we distinguish customer groups by their budget distributions, as the retailer would like to attract different types of customers to the store and decide which products to include in his assortment for profit maximization. Fig. 1 demonstrates optimal purchase baskets of different customers with isoelastic utility: on one side, risk-neutral customers choose the cheapest product among all the products and do not reduce the risk by diversification (Fig. 1a); on the other side, risk-averse customers diversify among the products to reduce their risk and prefer higher quantities of either standard or variable products.

Fig. 1(b) shows the optimal basket of a loyal customer with parameters $\gamma = 0.05$ and $\delta = 0.9$. Fig. 1(c) demonstrates the optimal basket of a non-loyal customer with parameters $\overline{\gamma} = 0.7$ and $\overline{\delta} = 0.4$. Compared to the article (Katsifou et al., 2014) and to the previous works in the field, our work accounts for optimal purchase quantities of customers by allowing them to buy more than one unit of each product under their budget constraint.

Next, based on the solution of the optimization problem (3), we can derive demands $d_j^s(\boldsymbol{s},\boldsymbol{v})$ and $d_j^v(\boldsymbol{s},\boldsymbol{v})$

$$d_j^s(\boldsymbol{s},\boldsymbol{v}) = \sum_{i\in\mathcal{L}} \mathbb{P}_{ir}\zeta_{ij}^s + \sum_{i\in\overline{\mathcal{L}}} \overline{\mathbb{P}}_{ir}\overline{\zeta}_{ij}^s \quad \text{and}$$

$$d_j^v(\boldsymbol{s},\boldsymbol{v}) = \sum_{i\in\mathcal{L}} \mathbb{P}_{ir}\xi_{ij}^v + \sum_{i\in\overline{\mathcal{L}}} \overline{\mathbb{P}}_{ir}\overline{\xi}_{ij}^v,$$

where $\mathbb{P}_{ir}$ and $\overline{\mathbb{P}}_{ir}$ are the probabilities to choose the retailer r for loyal and non-loyal customers correspondingly and we denote the optimal purchase quantities for non-loyal customers by $\overline{\zeta}_{ij}^s$ and $\overline{\xi}_{ij}^v$. The probabilities $\mathbb{P}_{ir}$ and $\overline{\mathbb{P}}_{ir}$ are modeled via a simplistic choice model (Anderson et al., 1992; Ben-Akiva & Lerman, 1994) with $\mathbb{P}_{ir} = 1 - F\left(-\alpha U_i^*\right)$ using the retailer r's market share $\alpha \in [0,1]$ and the standard Gumbel distribution $F(\cdot)$ for unobserved customers' preferences ϵ (Kök & Fisher, 2007; Smith & Agrawal, 2000; Smith, 2009). Note that the probability $\mathbb{P}_{ir}$ is an increasing function of U_i^* and does not depend on the total number of retailers. Further, the store choice probability can be bounded from below and from above by $\int_{\epsilon=0}^{\infty} f(\epsilon)d\epsilon \le \mathbb{P}_{ir} \le \int_{\epsilon=-U_i^*}^{\infty} f(\epsilon)d\epsilon$, where neither the upper nor the lower bound depends on the market share α, which makes them easily usable, especially in case of absence of any information about the market share or if the retailer expects it to change due to the introduction of variable products. Differently, one could model the probability that a customer i selects the retailer r among known number of stores R via the MNL model, assuming independence between customers' preferences, or via nested choice models, taking correlations between these preferences into account (Anderson et al., 1992; Ben-Akiva & Lerman, 1994; Flores, Berbeglia, & Hentenryck, 2019; Katsifou et al., 2014; Smith, 2009).

Note that the demands $d_j^s(\boldsymbol{s},\boldsymbol{v})$ and $d_j^v(\boldsymbol{s},\boldsymbol{v})$ are random variables, where randomness comes from the uncertainty in customer i's budget $B_i \sim H_i(x)$. Means of the demands are, therefore, to be computed as $\mu_j^s(\boldsymbol{s},\boldsymbol{v}) = \mathbb{E}(d_j^s(\boldsymbol{s},\boldsymbol{v}))$ and $\mu_j^v(\boldsymbol{s},\boldsymbol{v}) = \mathbb{E}(d_j^v(\boldsymbol{s},\boldsymbol{v}))$. Further, we use Gaussian type demand distributions with means μ_j^s and μ_j^v and standard deviations $\sqrt{\mu_j^s}$ and $\sqrt{\mu_j^v}$ correspondingly as an approximation to the Poisson distribution (van Ryzin, 1999).

3.3. Optimization model

Gathering the profit functions (1) and (2) for each type of product, we can now formulate our optimization model, aiming to maximize the retailer's expected profit per time unit.

The decision variables are the binary vectors $\boldsymbol{s}$ and $\boldsymbol{v}$ representing the standard and variable assortment decisions, and the inventory levels Q_j^s, $j \in \mathcal{N}_s$ and Q_j^v, $j \in \mathcal{N}_v$. The maximization of the expected profit is subject to a joint shelf space constraint. The volume of a standard (respectively variable) product j is denoted by l_j^s (respectively by l_j^v), while the available shelf space is denoted by C.

Here, for simplicity, we assume that the store has no backroom (like in Aldi) but C could also represent the combined space in shelves and in the backroom if one exists, allowing for regular replenishments of the store from the backroom. The aforementioned description leads to the following problem:

$$\begin{aligned}
&\max_{\boldsymbol{s},\boldsymbol{v},Q_j^s,Q_j^v\,\forall j} \left[\textstyle\sum_{j\in N_s} S_j\mathbb{E}\left(\frac{\Pi_j^s}{T_j^s}\right) + \sum_{j\in N_v} V_j\mathbb{E}\left(\frac{\Pi_j^v}{T_j^v}\right)\right],\\
&\qquad \text{subject to } \textstyle\sum_{j\in N_s} S_jQ_j^sl_j^s + \sum_{j\in N_v} V_jQ_j^vl_j^v \le C,\\
&\qquad \boldsymbol{s} = (S_1,\dots,S_{N_s}),\ S_j \in \{0,1\},\ \forall j \in \mathcal{N}_s,\\
&\qquad Q_j^s \ge S_jQ_{j,\min}^s,\ Q_j^s \le S_j\tfrac{C}{l_j^s},\ \forall j \in \mathcal{N}_s,\\
&\qquad \boldsymbol{v} = (V_1,\dots,V_{N_v}),\ V_j \in \{0,1\},\ \forall j \in \mathcal{N}_v,\\
&\qquad Q_j^v \ge V_jQ_{j,\min}^v,\ Q_j^v \le V_j\tfrac{C}{l_j^v},\ \forall j \in \mathcal{N}_v,
\end{aligned} \tag{6}$$

where $Q_{j,\min}^s \ge 1$ and $Q_{j,\min}^v \ge 1$ are minimal order quantities for standard $j \in \mathcal{N}_s$ and variable $j \in \mathcal{N}_v$ products. Note that constraints $Q_j^s \ge S_jQ_{j,\min}^s$ and $Q_j^v \ge V_jQ_{j,\min}^v$ prevent retailers ordering zero quantities of products; otherwise, it would be possible to artificially increase the demand by setting $S_j = 1$ or $V_j = 1$ with quantity $Q_j = 0$ for some product j. This would not necessarily strongly influence the total cost, which would depend on the setup cost and on the loss-of-goodwill. To avoid this, we suppose that order quantities for products in the assortment should be greater than $Q_{j,\min}^s$ and $Q_{j,\min}^v$ $\forall j$ with minimal order quantities being greater than one item. To prevent retailers ordering non-zero quantity of some product j while setting $S_j = 0$ or $V_j = 0$, we impose the following constraints: $Q^s_j \le S_j \frac{C}{l^s_j}$ and $Q^v_j \le V_j \frac{C}{l^v_j}$. Note that these constraints do not reduce the feasible set due to the available capacity.

Next, computing the profit of a standard or a variable product j (Eqs. (1) and (2)) requires the review period (T^s_j or T^v_j). Setting the review periods to $T^s_j = \frac{Q^s_j}{\mu^s_j}$, $\forall j \in \mathcal{N}_s$ and $T^v_j = \frac{Q^v_j}{\mu^v_j}$, $\forall j \in \mathcal{N}_v$ would make the optimization problem (6) not convex-concave in the retailer's order quantities and in mean demands, which depend on the customer's choice model. Furthermore, for marketing reasons, the review periods often need to be decided beforehand. Therefore, we would like to use such estimates for the periods T^s_j or T^v_j, that are independent of order quantities or mean demands.

Knowing that $Q^s_j \le \frac{C}{l^s_j}$, $\forall j \in \mathcal{N}_s$ and $Q^v_j \le \frac{C}{l^v_j}$, $\forall j \in \mathcal{N}_v$, we use $T^s_j = \frac{2k^s}{l^s_j \hat{\mu}^s_j}$, $\forall j \in \mathcal{N}_s$ and $T^v_j = \frac{2k^v}{l^v_j \hat{\mu}^v_j}$, $\forall j \in \mathcal{N}_v$ for the approximations of the review periods in our numerical tests with $T^s_j > T^v_j$, k^s and k^v being some constants and $\hat{\mu}^s_j$, $\hat{\mu}^v_j$ being the upper bounds of mean demands introduced later in Section 4.3.

The assortment planning problem we study is a *bilevel optimization problem*, where the retailer needs to solve the constrained mixed integer non-linear problem (6) with means μ^s_j and μ^v_j and standard deviations $\sqrt{\mu^s_j}$ and $\sqrt{\mu^v_j}$ dependent on the solution of the customer's concave optimization problem (3), which is dependent on the assortment vectors $\boldsymbol{s}$ and $\boldsymbol{v}$. The mixed integer non-linear problem (6) bears similarities with the knapsack problem (Garey & Johnson, 1979), which is NP-hard. However, our assortment problem is clearly more complex as the profit associated with each product is a non-linear function of the inventory level, and is dependent on other selected products. Moreover, the retailer's product selection depends on the continuous distribution of the customers' budget. This significantly complicates the estimation of mean demands, making random sampling very inefficient and inaccurate.

4. Solution method

4.1. Optimal quantization of probability distributions

The retailer r solves the optimization problem (6) under uncertain demands for standard and variable products D^s_j, $j \in \mathcal{N}_s$ and D^v_j $j \in \mathcal{N}_v$ and uncertain customer i's budget B_i. These random variables are given by their continuous distribution functions: we use normal distributions, G^s_j and G^v_j, with means $T^s_j \mu^s_j$ and $T^v_j \mu^v_j$ and standard deviations $\sqrt{T^s_j \mu^s_j}$ and $\sqrt{T^v_j \mu^v_j}$ for total demands D^s_j and D^v_j correspondingly; for customer i's budget, we use the lognormal distribution H_i with mean and variance described in (5).

In order to solve the optimization problem (6), the distribution functions need to be discretized, i.e., we need to find discrete distributions, sitting on N points, which approximate our continuous distributions at best. For this, we use *optimal quantization* (see Monge, 1781; Rachev & Rüschendorf, 1998; Villani, 2008) rather than random (i.e., Monte Carlo) sampling with the aim to receive an optimal rather than a random solution, minimizing a distance measure. We use Kantorovich-Wasserstein distance (see Kantorovich, 1942; Villani, 2008) between probability measures, as the measure of goodness of the approximation (Appendix 1.3). We denote discrete approximations of probability measures corresponding to continuous distributions, G^s_j for standard $j \in \mathcal{N}_s$ and G^v_j for variable products $j \in \mathcal{N}_v$, by $\widetilde{P}^s_j = \sum_{i=1}^N \widetilde{p}^s_{ij} \delta_{\widetilde{z}^s_{ij}}$ and $\widetilde{P}^v_j = \sum_{i=1}^N \widetilde{p}^v_{ij} \delta_{\widetilde{z}^v_{ij}}$ respectively, with corresponding vectors of optimal supporting points $\widetilde{\boldsymbol{z}}^{\boldsymbol{s}}_{\boldsymbol{j}}$, $\widetilde{\boldsymbol{z}}^{\boldsymbol{v}}_{\boldsymbol{j}}$ and probabilities $\widetilde{\boldsymbol{p}}^{\boldsymbol{s}}_{\boldsymbol{j}}$, $\widetilde{\boldsymbol{p}}^{\boldsymbol{v}}_{\boldsymbol{j}}$. Further, discrete approximations of budget distributions for loyal and non-loyal customers sit correspondingly on optimal supporting points $(B_1, B_2, \dots, B_N)$ with probabilities $(b_1, b_2, \dots, b_N)$ and on optimal supporting points $(\overline{B}_1, \overline{B}_2, \dots, \overline{B}_N)$ with probabilities $(\overline{b}_1, \overline{b}_2, \dots, \overline{b}_N)$. Next, we reformulate the optimization problem (6) in a way suitable for its numerical solution, where all the discrete approximations are sitting on N optimal supporting points. Clearly, the larger the number N of optimal supporting points the finer the approximation.

4.2. Problem reformulation

Now, we reformulate the optimization problem (6) for a numerical solution. For this, we introduce additional vectors $\widetilde{\boldsymbol{q}}^s_j = (\widetilde{q}^s_{1j}, \widetilde{q}^s_{2j}, \dots, \widetilde{q}^s_{Nj})$ and $\widetilde{\boldsymbol{q}}^v_j = (\widetilde{q}^v_{1j}, \widetilde{q}^v_{2j}, \dots, \widetilde{q}^v_{Nj})$, whose elements should satisfy the following properties for fixed Q^s_j and Q^v_j and for all i:

$$\widetilde{q}^s_{ij} = 0 \text{ if } \widetilde{z}^s_{ij} - Q^s_j < 0, \text{ otherwise } \widetilde{q}^s_{ij} = \widetilde{z}^s_{ij} - Q^s_j \text{ if } \widetilde{z}^s_{ij} - Q^s_j \ge 0,$$
$$\widetilde{q}^v_{ij} = 0 \text{ if } \widetilde{z}^v_{ij} - Q^v_j < 0, \text{ otherwise } \widetilde{q}^v_{ij} = \widetilde{z}^v_{ij} - Q^v_j \text{ if } \widetilde{z}^v_{ij} - Q^v_j \ge 0. \quad (7)$$

The vectors $\widetilde{\boldsymbol{q}}^s_j$ and $\widetilde{\boldsymbol{q}}^v_j$ are introduced in order to approximate integrals $\int_{Q^s_j}^{\infty} (y - Q^s_j) g^s_j(y) dy$, $\int_{Q^v_j}^{\infty} (y - Q^v_j) g^v_j(y) dy$ and $\int_0^{Q^v_j} (Q^v_j - y) g^v_j(y) dy$ (see Appendix 1.2 for details). To satisfy Eq. (7), decision variables $\widetilde{\boldsymbol{q}}^s_j$ and $\widetilde{\boldsymbol{q}}^v_j$ need to attain their minimal values in the feasible set. This is guaranteed as the approximated integrals enter the maximization problem (6) in the cost part.

Therefore, the retailer's optimization problem yields:

$$
\begin{aligned}
&\max_{\boldsymbol{s},\boldsymbol{v},Q^s_j,Q^v_j,\widetilde{\boldsymbol{q}}^s_j,\widetilde{\boldsymbol{q}}^v_j \,\forall j} \left[\sum_{j\in\mathcal{N}_s} S_j \mathbb{E}\left(\frac{\Pi^s_j}{T^s_j}\right) + \sum_{j\in\mathcal{N}_v} V_j \mathbb{E}\left(\frac{\Pi^v_j}{T^v_j}\right) \right], \\
&\qquad \text{subject to } \sum_{j\in N_s} Q^s_j l^s_j + \sum_{j\in N_v} Q^v_j l^v_j \le C, \\
&\qquad \widetilde{\boldsymbol{z}}^s_j(\boldsymbol{s},\boldsymbol{v}) - Q^s_j \boldsymbol{e} \le \widetilde{\boldsymbol{q}}^s_j, \\
&\qquad \widetilde{\boldsymbol{q}}^s_j \ge 0, \ \forall j \in \mathcal{N}_s, \quad \widetilde{\boldsymbol{z}}^v_j(\boldsymbol{s},\boldsymbol{v}) - Q^v_j \boldsymbol{e} \le \widetilde{\boldsymbol{q}}^v_j, \\
&\qquad \widetilde{\boldsymbol{q}}^v_j \ge 0, \ \forall j \in \mathcal{N}_v, \\
&\qquad \mathbb{E}\left(\frac{\Pi^s_j}{T^s_j}\right) = m^s_j \mu^s_j(\boldsymbol{s},\boldsymbol{v}) \\
&\qquad - \left[\frac{K^s_j}{T^s_j} + \frac{h^s_j}{T^s_j}\Big(Q^s_j - T^s_j \mu^s_j(\boldsymbol{s},\boldsymbol{v})\Big) + \frac{cu^s_j}{T^s_j} \widetilde{\boldsymbol{p}}^s_j(\boldsymbol{s},\boldsymbol{v}) \cdot \widetilde{\boldsymbol{q}}^s_j\right], \\
&\qquad \mathbb{E}\left(\frac{\Pi^v_j}{T^v_j}\right) = m^v_j \mu^v_j(\boldsymbol{s},\boldsymbol{v}) \\
&\qquad - \left[\frac{K^v_j}{T^v_j} + \frac{c^v_j + co^v_j}{T^v_j}\Big(Q^v_j - T^v_j \mu^v_j(\boldsymbol{s},\boldsymbol{v})\Big) \right. \\
&\qquad \left. + \frac{cu^v_j + co^v_j}{T^v_j} \widetilde{\boldsymbol{p}}^v_j(\boldsymbol{s},\boldsymbol{v}) \cdot \widetilde{\boldsymbol{q}}^v_j\right], \\
&\qquad Q^s_j \ge S_j Q^s_{j,\min}, \ Q^s_j \le S_j \frac{C}{l^s_j}, \ \forall j \in \mathcal{N}_s, \\
&\qquad Q^v_j \ge V_j Q^v_{j,\min}, \ Q^v_j \le V_j \frac{C}{l^v_j}, \ \forall j \in \mathcal{N}_v, \\
&\qquad \boldsymbol{s} = (S_1, \dots, S_{N_s}), \ S_j \in \{0,1\}, \ \forall j \in \mathcal{N}_s, \\
&\qquad \boldsymbol{v} = (V_1, \dots, V_{N_v}), \ V_j \in \{0,1\}, \ \forall j \in \mathcal{N}_v,
\end{aligned}
\quad (8)
$$

where mean demands for L loyal and $\overline{L}$ non-loyal customers with N budget quantizers are computed as

$\mu_j^s(\boldsymbol{s},\boldsymbol{v}) = L\sum_{i=1}^N \mathbb{P}_{ir} b_i \zeta_{ij}^s + \bar{L}\sum_{i=1}^N \overline{\mathbb{P}}_{ir}\bar{b}_i \overline{\zeta}_{ij}^s \quad \forall j \in \mathcal{N}_s$ and $\mu_j^v(\boldsymbol{s},\boldsymbol{v}) = L\sum_{i=1}^N \mathbb{P}_{ir} b_i \xi_{ij}^v + \bar{L}\sum_{i=1}^N \overline{\mathbb{P}}_{ir}\bar{b}_i \overline{\xi}_{ij}^v \ \forall j \in \mathcal{N}_v$ correspondingly.

Note that the customer's optimization problem (3) uniquely defines optimal purchase quantities, optimal utilities and the store choice probability, given the type of customer (loyal or non-loyal) and his budget quantizer B_i or $\bar{B}_i$. This is due to the fact that the solution of the optimization problem (3) under (4) and (5) depends solely on the customer's budget and the customer's type. The above equations for mean demands, therefore, hold. Furthermore, one could adapt the approach for the estimation of mean demands $\mu_j^s(\boldsymbol{s},\boldsymbol{v})$ and $\mu_j^v(\boldsymbol{s},\boldsymbol{v})$ for the case when the number of customers is unknown: instead of directly choosing L and $\bar{L}$ as the number of loyal and non-loyal customers, one could fix them to be equal to one, but account for the true number of customers per time unit by changing time units, risk aversion and the budget distribution parameters (4), (5), that are possible to estimate based on the retailer's data. We, however, proceed with the equations above.

Optimization problem (8) is easily solved numerically for $Q_j^s,\ \forall j \in \mathcal{N}_s,\ Q_j^v,\ \forall j \in \mathcal{N}_v$ under fixed assortment vectors $\boldsymbol{s}$ and $\boldsymbol{v}$. Clearly, with $\boldsymbol{s}$ and $\boldsymbol{v}$ being the decision variables, the exact solution can be found only with complete enumeration. However, the computational time for complete enumeration (having to examine $(2^{N_s+N_v}-1)$ combinations for a set of N_s+N_v candidate products) quickly becomes prohibitive for large-scale problems. Thus, we propose several lower and upper bounds, with the aim of finding a good solution for large-size problems. In the following we present upper and lower bounds of the problem (8) and, afterwards, we generate test data for the numerical solution.

4.3. Upper bounds

If the expected demands $\mu_j^s(\boldsymbol{s},\boldsymbol{v}),\ \mu_j^v(\boldsymbol{s},\boldsymbol{v})$, optimal quantizers $\widetilde{\boldsymbol{z}}_j^s(\boldsymbol{s},\boldsymbol{v}),\ \widetilde{\boldsymbol{z}}_j^v(\boldsymbol{s},\boldsymbol{v})$ and corresponding probabilities $\widetilde{\boldsymbol{p}}_j^s(\boldsymbol{s},\boldsymbol{v}),\ \widetilde{\boldsymbol{p}}_j^v(\boldsymbol{s},\boldsymbol{v})$ were independent of assortment vectors $\boldsymbol{s}$ and $\boldsymbol{v}$ in the problem (8) $\forall j \in \mathcal{N}_s,\ \forall j \in \mathcal{N}_v$, being fixed at some levels $\mu_j^s,\ \mu_j^v,\ \widetilde{\boldsymbol{z}}_j^s,\ \widetilde{\boldsymbol{z}}_j^v,\ \widetilde{\boldsymbol{p}}_j^s,\ \widetilde{\boldsymbol{p}}_j^v$, we would derive continuous relaxations, such as *linear* and *SDP relaxations*, to bound the optimal value of the problem (8) from above. However, all these functions are dependent on assortment vectors $\boldsymbol{s}$ and $\boldsymbol{v}$ in our case. Therefore, to impose the upper bound on the problem (8), we first impose upper bounds on $\mu_j^s(\boldsymbol{s},\boldsymbol{v}),\ \forall j \in \mathcal{N}_s$ and $\mu_j^v(\boldsymbol{s},\boldsymbol{v}),\ \forall j \in \mathcal{N}_v$, i.e., we find $\hat{\mu}_j^s$ and $\hat{\mu}_j^v$, independent of $\boldsymbol{s}$ and $\boldsymbol{v}$, such that $0 \le \mu_j^s(\boldsymbol{s},\boldsymbol{v}) \le \hat{\mu}_j^s$ and $0 \le \mu_j^v(\boldsymbol{s},\boldsymbol{v}) \le \hat{\mu}_j^v,\ \forall j,\boldsymbol{s},\boldsymbol{v}$, and we include the mean functions to the set of decision variables in linear and SDP relaxations.

There are different methods for the estimation of upper bounds $\hat{\mu}_j^s$ and $\hat{\mu}_j^v$. It would be possible to compute the bounds based on the data on demand. However, as we do not have access to the retailer's data, we propose a simple approach to compute the expected demand upper bounds based on our model.

First of all, consider the upper bound $\mathbb{P}_{ir} \le 1 - F(-U_i^*)$, which depends only on the customer i's optimal utility U_i^*. The higher the utility, the larger the probability that the customer chooses the retailer r. Note that the utility is at absolute maximum if S_j and V_j are equal to 1 $\forall j$ in the customer's model (3), i.e., all products are available at the retailer and customers assume the ability to purchase as much as they wish under their budget constraints. We denote the corresponding upper bound on the store choice probability by $\mathbb{W}_{ir}$ for a loyal and by $\overline{\mathbb{W}}_{ir}$ for a non-loyal customer i. The maximal purchase quantity for each product is just the customer's budget divided by the price of the product. The upper bounds of the expected demands are, therefore, equal to $\hat{\mu}_j^s = L\sum_{i=1}^N \mathbb{W}_{ir} b_i \frac{B_i}{p_j^s} + \bar{L}\sum_{i=1}^N \overline{\mathbb{W}}_{ir}\bar{b}_i \frac{\bar{B}_i}{p_j^s} \quad \forall j \in \mathcal{N}_s$ and $\hat{\mu}_j^v = L\sum_{i=1}^N \mathbb{W}_{ir} b_i \frac{B_i}{p_j^v} + \bar{L}\sum_{i=1}^N \overline{\mathbb{W}}_{ir}\bar{b}_i \frac{\bar{B}_i}{p_j^v}\ \forall j \in \mathcal{N}_v$.

In linear and SDP relaxations, we suppose that mean demands for standard and variable products are decision variables belonging to intervals $[0,\ \hat{\mu}_j^s]$ and $[0,\ \hat{\mu}_j^v]$ correspondingly. Also, we notice that the optimal quantizers $\widetilde{\boldsymbol{z}}_j^s(\boldsymbol{s},\boldsymbol{v}),\ \widetilde{\boldsymbol{z}}_j^v(\boldsymbol{s},\boldsymbol{v})$ (as well as the probabilities $\widetilde{\boldsymbol{p}}_j^s(\boldsymbol{s},\boldsymbol{v})$ and $\widetilde{\boldsymbol{p}}_j^v(\boldsymbol{s},\boldsymbol{v})$) depend on assortment vectors $\boldsymbol{s}$ and $\boldsymbol{v}$ only through the mean functions $\mu_j^s(\boldsymbol{s},\boldsymbol{v})$, $\mu_j^v(\boldsymbol{s},\boldsymbol{v})$ as $\widetilde{\boldsymbol{z}}_j^s(\boldsymbol{s},\boldsymbol{v}) = T_j^s\mu_j^s(\boldsymbol{s},\boldsymbol{v})\boldsymbol{e} + \sqrt{T_j^s\mu_j^s(\boldsymbol{s},\boldsymbol{v})}\widetilde{\boldsymbol{z}}_j^s(0)$ and $\widetilde{\boldsymbol{z}}_j^v(\boldsymbol{s},\boldsymbol{v}) = T_j^v\mu_j^v(\boldsymbol{s},\boldsymbol{v})\boldsymbol{e} + \sqrt{T_j^v\mu_j^v(\boldsymbol{s},\boldsymbol{v})}\widetilde{\boldsymbol{z}}_j^v(0)$, where $\widetilde{\boldsymbol{z}}_j^s(0)$ and $\widetilde{\boldsymbol{z}}_j^v(0)$ are optimal quantizers of the standard Gaussian distribution with corresponding probability vectors $\widetilde{\boldsymbol{p}}_j^s(0)$ and $\widetilde{\boldsymbol{p}}_j^v(0)$. For simplicity, we use the optimal weights $\widetilde{\boldsymbol{p}}_j^s(0)$ and $\widetilde{\boldsymbol{p}}_j^v(0)$ of the standard Gaussian distribution instead of $\widetilde{\boldsymbol{p}}_j^s(\boldsymbol{s},\boldsymbol{v})$ and $\widetilde{\boldsymbol{p}}_j^v(\boldsymbol{s},\boldsymbol{v})$ in all our upper and lower bounds (alternatively, one could use uniform probabilities): the higher the number of quantizers the better the approximation.

To guarantee the upper bound and to have only linear constraints on $\widetilde{\boldsymbol{q}}_j^s,\ \widetilde{\boldsymbol{q}}_j^v$ in the relaxation, we use the linear approximation of the square root function going through points (0, 0) and $(T_j^s\hat{\mu}_j^s, \sqrt{T_j^s\hat{\mu}_j^s})$ for standard or $(T_j^v\hat{\mu}_j^v, \sqrt{T_j^v\hat{\mu}_j^v})$ for variable products. Using the linear approximation for $\widetilde{\boldsymbol{z}}_j^s(\boldsymbol{s},\boldsymbol{v})$ and $\widetilde{\boldsymbol{z}}_j^v(\boldsymbol{s},\boldsymbol{v})$ in the relaxations, one guarantees that $(\widetilde{\boldsymbol{p}}_j^s(0), \widetilde{\boldsymbol{q}}_j^s)$ and $(\widetilde{\boldsymbol{p}}_j^v(0), \widetilde{\boldsymbol{q}}_j^v)$ will attain lower values and, therefore, the stock-out cost in the problem (8) will drop, leading to the upper bound on the profit (see Lemma 1.1 in Appendix 1.5 for the proof).

Linear relaxation with decision variables $X_j^s = S_j\mu_j^s$ and $X_j^v = V_j\mu_j^v,\ Y_j^s = S_jQ_j^s$ and $Y_j^v = V_jQ_j^v,\ Z_j^s = S_j\cdot(\widetilde{\boldsymbol{p}}_j^s(0), \widetilde{\boldsymbol{q}}_j^s)$ and $Z_j^v = V_j\cdot(\widetilde{\boldsymbol{p}}_j^v(0), \widetilde{\boldsymbol{q}}_j^v)$ can be written in the following standard form introduced by Fortet (1960):

$$
\begin{aligned}
&\max \Bigg[\sum_{j\in\mathcal{N}_s} \Bigg\{ (m_j^s + h_j^s)X_j^s - \frac{K_j^s}{T_j^s}S_j - \frac{h_j^s}{T_j^s}Y_j^s - \frac{cu_j^s}{T_j^s}Z_j^s \Bigg\} + \sum_{j\in\mathcal{N}_v} \\
&\quad \times \Bigg\{ (m_j^v + c_j^v + co_j^v)X_j^v - \frac{K_j^v}{T_j^v}V_j - \frac{c_j^v + co_j^v}{T_j^v}Y_j^v - \frac{cu_j^v + co_j^v}{T_j^v}Z_j^v \Bigg]\Bigg\}, \\
&\text{subject to } 0 \le S_j \le 1, \quad 0 \le \mu_j^s \le \hat{\mu}_j^s,\ \forall j \in \mathcal{N}_s, \\
&0 \le V_j \le 1,\ 0 \le \mu_j^v \le \hat{\mu}_j^v, \quad \forall j \in \mathcal{N}_v, \\
&T_j^s\mu_j^s\boldsymbol{e} + \sqrt{\frac{T_j^s}{\hat{\mu}_j^s}}\mu_j^s\widetilde{\boldsymbol{z}}_j^s(0) - Q_j^s\boldsymbol{e} \le \widetilde{\boldsymbol{q}}_j^s,\ \widetilde{\boldsymbol{q}}_j^s \ge 0,\ \forall j \in \mathcal{N}_s, \\
&T_j^v\mu_j^v\boldsymbol{e} + \sqrt{\frac{T_j^v}{\hat{\mu}_j^v}}\mu_j^v\widetilde{\boldsymbol{z}}_j^v(0) - Q_j^v\boldsymbol{e} \le \widetilde{\boldsymbol{q}}_j^v,\ \widetilde{\boldsymbol{q}}_j^v \ge 0,\ \forall j \in \mathcal{N}_v, \\
&\sum_{j\in N_s} Y_j^s l_j^s + \sum_{j\in N_v} Y_j^v l_j^v \le C, \quad Q_j^s \le S_j\frac{C}{l_j^s},\ \forall j \in \mathcal{N}_s, \\
&Q_j^v \le V_j\frac{C}{l_j^v},\ \forall j \in \mathcal{N}_v, \\
&0 \le X_j^s \le \mu_j^s, \quad X_j^s \ge \mu_j^s + S_j - 1,\ \forall j \in \mathcal{N}_s, \\
&0 \le X_j^v \le \mu_j^v, \quad X_j^v \ge \mu_j^v + V_j - 1,\ \forall j \in \mathcal{N}_v, \\
&S_jQ_{j,\min}^s \le Y_j^s \le Q_j^s,\ \forall j \in \mathcal{N}_s, \quad V_jQ_{j,\min}^v \le Y_j^v \le Q_j^v,\ \forall j \in \mathcal{N}_v, \\
&Y_j^s \ge Q_j^s + S_j - 1,\ \forall j \in \mathcal{N}_s, \quad Y_j^v \ge Q_j^v + V_j - 1,\ \forall j \in \mathcal{N}_v, \\
&0 \le Z_j^s \le (\widetilde{\boldsymbol{p}}_j^s(0), \widetilde{\boldsymbol{q}}_j^s),\ \forall j \in \mathcal{N}_s, \\
&0 \le Z_j^v \le (\widetilde{\boldsymbol{p}}_j^v(0), \widetilde{\boldsymbol{q}}_j^v),\ \forall j \in \mathcal{N}_v,
\end{aligned}
$$

$$Z_j^s \geq (\widetilde{\boldsymbol{p}}_j^s(0), \widetilde{\boldsymbol{q}}_j^s) + (S_j - 1)\hat{\mu}_j^s,\ \forall j \in \mathcal{N}_s,$$
$$Z_j^v \geq (\widetilde{\boldsymbol{p}}_j^v(0), \widetilde{\boldsymbol{q}}_j^v) + \ (V_j - 1)\hat{\mu}_j^v,\ \forall j \in \mathcal{N}_v. \tag{9}$$

Similar linear relaxation is used in Gaivoronski et al. (2011) for the quadratic Knapsack problem with probability constraints. As in Gaivoronski et al. (2011), this relaxation provides a fast way to have an upper bound on the optimization problem (8). Relaxation (9) does not necessarily result in a very tight upper bound. Therefore, we also construct a SDP relaxation of the problem (8). For this, we first rewrite the problem (8) as a zero-one QCQP (quadratically constrained quadratic problem):

$$\begin{aligned} &\max_{\boldsymbol{x}} \boldsymbol{x}^T \hat{\boldsymbol{\Pi}} \boldsymbol{x} \\ &\quad \text{subject to } \boldsymbol{x} \geq 0, \\ &\quad \textstyle\sum_{i=1}^K \omega_i^k x_i \leq \alpha_k,\ \forall k = 1, \dots, K, \\ &\quad x_i \in \{0,\ 1\},\ \forall i = 1, \dots, N_s, N_s(3+N)+1, \dots, N_s(3+N)+N_v, \end{aligned} \tag{10}$$

where $K = (3+N)(N_s+N_v)$; $\hat{\boldsymbol{\Pi}}$ is an appropriate $K \times K$ sparse matrix; $\boldsymbol{\omega}^k = (\omega_1^k, \dots, \omega_K^k)$ are appropriate K-vectors constructed for the objective function and all linear constraints in the problem (8) with $\alpha_k, \forall k = 1, \dots, K$ being the corresponding constants; decision vector $\boldsymbol{x}$ is defined by $\boldsymbol{x}^T = (\boldsymbol{s}, Q_1^s, \dots, Q_{N_s}^s, \mu_1^s, \dots, \mu_{N_s}^s, \widetilde{\boldsymbol{q}}_1^s, \dots, \widetilde{\boldsymbol{q}}_{N_s}^s, \boldsymbol{v}, Q_1^v, \dots, Q_{N_v}^v, \mu_1^v, \dots, \mu_{N_v}^v, \widetilde{\boldsymbol{q}}_1^v, \dots, \widetilde{\boldsymbol{q}}_{N_v}^v)$.

Further, let $\boldsymbol{X}$ be the matrix $\begin{bmatrix} \boldsymbol{x}^T\boldsymbol{x} & \boldsymbol{x}^T \\ \boldsymbol{x} & 1 \end{bmatrix}$. We will use the modified matrix $\tilde{\pi} = \begin{bmatrix} \hat{\boldsymbol{\Pi}} & 0 \\ 0 & 0 \end{bmatrix}$ with $\boldsymbol{W}_k$ being the appropriate matrices constructed based on constraints of (10). Using these notations, we can easily build a SDP relaxation of (10). In relaxation (11), we tighten the constraints so that the binary variables take values as close to 0 and 1 as possible. This is done to strengthen the approximation shown to be weak in the studies of Helmberg, Rendl, and Weismantel (2000) for the quadratic knapsack problem, Rendl and Sotirov (2003) for the quadratic assignment problem and Gaivoronski et al. (2011) for the quadratic knapsack problem with probability constraints. Similar to the work of Gaivoronski et al. (2011), we use *Sherali-Adams constraints* proposed in Sherali and Adams (1990) and Sherali and Adams (1994), which are generated by the multiplication of each constraint by the binary variable it contains: this limits the space of solutions pushing binary variables to take values closer to 0 and 1. The only constraints containing binary variables S_j or V_j are $Q_j^s \geq S_j Q_{j,\min}^s$, $Q_j^s \leq S_j \frac{C}{l_j^s}$, $\forall j \in \mathcal{N}_s$ and $Q_j^v \geq V_j Q_{j,\min}^v$, $Q_j^v \leq V_j \frac{C}{l_j^v}$, $\forall j \in \mathcal{N}_v$ and none of the constraints contain several binary variables at once. This means an increase of $N_s + N_v$ in the number of linear constraints, as only $N_s + N_v$ constraints are multiplied by S_j, $1 - S_j$ or V_j, $1 - V_j$.

Similar to the article (Gaivoronski et al., 2011), we do not construct a sequence of relaxations proposed in Sherali and Adams (1990, 1994) but limit ourselves to the first relaxation in the sequence. Though the sequence of relaxations in Sherali and Adams (1990, 1994) would lead to the integer polytope, building each of the relaxations would soon become computationally expensive. Furthermore, the obtained result would still be an upper bound on the problem (8) due to the fact that the mean demands μ_j^s and μ_j^v are included in the set of decision variables in our relaxations and that we use the approximations (14) in Appendix 1.5. The semidefinite (SDP) relaxation with Sherali-Adams constraints yields

$$\begin{aligned} \max_{\boldsymbol{X},\boldsymbol{x}}\ &\text{trace}(\widetilde{\boldsymbol{\Pi}} \bullet \boldsymbol{X}) \\ &\text{subject to } \ \boldsymbol{x} \geq 0,\ \boldsymbol{X} \succ 0, \\ &\text{trace}(\widetilde{\boldsymbol{W}}_k^{j(k)} \bullet \boldsymbol{X}) \leq 0,\ \forall k = 1, \dots, 2N_s, (3+N)N_s + 1, \dots, \\ &\quad (3+N)N_s + 2N_v, \\ &\text{trace}(\hat{\boldsymbol{W}}_k^{j(k)} \bullet \boldsymbol{X}) \leq \alpha_k,\ \forall k = 1, \dots, 2N_s, (3+N)N_s + 1, \dots, \\ &\quad (3+N)N_s + 2N_v \\ &\text{trace}(\boldsymbol{W}_k \bullet \boldsymbol{X}) \leq \alpha_k,\ \forall k = 2N_s + 1, \dots, (2+N)N_s, \\ &\text{trace}(\boldsymbol{W}_k \bullet \boldsymbol{X}) \leq \alpha_k,\ \forall k = (2+N)N_s + 2N_v + 1, \dots, \\ &\quad (2+N)(N_s + N_v) \\ &\text{diag}\Big((x_1, \dots, x_{N_s})^T (x_1, \dots, x_{N_s})\Big) = (x_1, \dots, x_{N_s})^T, \\ &\text{diag}\Big((x_{(3+N)N_s+1}, \dots, x_{(3+N)N_s+N_v})^T (x_{(3+N)N_s+1}, \dots, \\ &\quad x_{(3+N)N_s+N_v})\Big) = \\ &= (x_{(3+N)N_s+1}, \dots, x_{(3+N)N_s+N_v})^T, \end{aligned} \tag{11}$$

where $\widetilde{\boldsymbol{W}}_k^{j(k)}$ and $\hat{\boldsymbol{W}}_k^{j(k)}$ are the matrices corresponding to Sherali-Adams constraints described above and the multiplication is by binary variables S_j (or V_j) and $1 - S_j$ (or $1 - V_j$), where j is uniquely defined by k.

Note that rounding optimal solutions of relaxations (9) and (11) would give some assortment vectors that could be used for the computation of simplistic lower bounds of the problem (8). Using the linear relaxation and rounding the solution, one would gain computational efficiency for very high-dimensional problems. Below in this section, we test the efficiency and accuracy of the proposed lower and upper bounds.

4.4. Lower bounds

It is quite easy to construct multiple lower bounds for the problem (8) by taking subsets of possible realizations of vectors $\boldsymbol{s}$ and $\boldsymbol{v}$ and, therefore, bounding the feasible set of the problem. The result does not necessarily find the optimal product allocation but is useful for the error bounding and, moreover, for the decision about the assortment (as the retailer can be confident that he would not be worse off).

We consider several lower bounds for the optimization problem (6):

Rank and optimize: Suppose that only one single product is available in the retailer's assortment. Thus, the assortment vectors $\boldsymbol{s}$ and $\boldsymbol{v}$ are fixed, containing only one single 1 in one of them. In this case, the objective function of the problem (6) degenerates to the profit resulted from the sales of the available product only, which is a concave function of the inventory level (it can be easily proven by taking the derivative of the objective function in (6)). We face a problem similar to the knapsack problem but with a concave profit function (instead of linear). One can maximize the profit and receive the quantity of the product, which is only optimal for these particular assortment vectors $\boldsymbol{s}$ and $\boldsymbol{v}$ but not for the initial problem.

However, instead of maximizing the expected profit function, one can try to maximize the expected profit to space ratio of the available product. This arises from the consideration that, due to the limited shelf space, the retailer has to select those products that bring profit, without filling too much space. A product j can be thus selected based on its expected profit to space ratio.

The profit to space ratio has previously been adopted by Lim, Rodrigues, and Zhang (2004) and Yang (2001) to solve the retail shelf space allocation problem as well as by Hansen and Heinsbroek (1979) in the joint problem of product selection and shelf space allocation. In our model, the expected profit to space ratio is given by $\mathbb{E}\big(\frac{\Pi_j^s}{l_j^s T_j^s Q_j^s}\big)$ for standard products $j \in \mathcal{N}_s$ and by $\mathbb{E}\big(\frac{\Pi_j^v}{l_j^v T_j^v Q_j^v}\big)$ for variable products $j \in \mathcal{N}_v$. The optimal slopes correspond to tangent points, i.e., to Q_j^s and Q_j^v such that

$$\int_{Q_j^s}^{\infty} y g_j^s(y) dy = \frac{(m_j^s + h_j^s)\mu_j^s T_j^s - K_j^s}{cu_j^s}, \quad \forall j \in \mathcal{N}_s,$$
$$\int_{Q_j^v}^{\infty} y g_j^v(y) dy = \frac{(m_j^v + c_j^v + co_j^v)\mu_j^v T_j^v - K_j^v}{cu_j^v + co_j^v}, \quad \forall j \in \mathcal{N}_v. \tag{12}$$

Sequentially going through the products, one can optimize the expected profit to space ratio for each of them and rank all the $N_s + N_v$ products in the assortment based on their optimal profit to space ratios in descending order. After, instead of $2^{N_s+N_v} - 1$ problems, one can solve $N_s + N_v$ optimization problems (8) with fixed assortment vectors $\boldsymbol{s}$ and $\boldsymbol{v}$, starting with the first-ranked product only and finishing with all products being included in the assortment. Choosing the maximal optimal value between these problems, one decides about the assortment and the optimal order quantities.

The number of iterations in the algorithm is, therefore, $2(N_s + N_v)$, where $N_s + N_v$ iterations are devoted to the ranking based on optimal profit to space ratios and other $N_s + N_v$ iterations are assigned to the solution of optimization problems (8) with fixed assortment vectors.

Tangent points Q_j^s and Q_j^v can be computed via different approximation algorithms based on implicit functions (12). If the points are estimated using optimal quantization (see Appendix1.2), we refer to the *Rank and Optimize* method as *TgRankOpt*. Differently, one can approximate optimal slopes by $\frac{(\Pi_j^s)^*}{l_j^s T_j^s (Q_j^s)^*}$, $\forall j \in \mathcal{N}_s$ and $\frac{(\Pi_j^v)^*}{l_j^v T_j^v (Q_j^v)^*}$, $\forall j \in \mathcal{N}_v$, where $(\Pi_j^s)^*$, $(\Pi_j^v)^*$, $(Q_j^s)^*$, $(Q_j^v)^*$ are optimal expected profits and optimal quantities corresponding to standard or variable products $j \forall j$ in the problem (8). In this case, we refer to the *Rank and Optimize* simply as *RankOpt*.

Rank and approximate: Analogically to the *Rank and Optimize* method, one ranks $N_s + N_v$ products in the assortment based on their expected profit to space ratios. Further, instead of solving $N_s + N_v$ optimization problems, one uses a sequential procedure to approximate the solution: (1) the product with the highest profit to space ratio is included first and the next selected product has the second highest profit to space ratio; (2) as the first included product has the highest marginal profit to space contribution, increasing its inventory level offers the best yield; thus, its inventory level is increased until it reaches equal marginal profit contribution with the second product; (3) the addition of a third product is then considered; in general, given the limited shelf space, the process continues as long as the shelf space is not exceeded: when the next product is included, the order quantities of the previously included products are updated so that the marginal profit to space contribution is equal for all of them; (4) the process terminates when m products are included in the assortment and the $(m+1)^{th}$ product would violate the space constraint.

Without constraints on minimal order quantities $Q_{j,\min}^s$ and $Q_{j,\min}^v$ $\forall j$, this lower bound is at best as tight as the one of *Rank and Optimize*, as the order quantities are suboptimal. With minimal lot sizes, one may be willing to set order quantities to the minimal amounts already at the ranking step for products with optimal tangent points below $Q_{j,\min}^s$ (or $Q_{j,\min}^v$). Further, computing suboptimal slopes at $Q_{j,\min}^s$ (or $Q_{j,\min}^v$) for these products, one influences the ranking of products in comparison to *RankOpt*. The number of iterations of the algorithm is less than or equal to $2(N_s + N_v)$, as the process terminates as soon as the capacity constraint is violated. Further, we refer to this method as *TgRankApprox*.

Bound and round: Using either LP relaxation (9) or SDP upper bound (11), one obtains optimal relaxed S_j^* and V_j^*, $\forall j$ from the interval $[0, 1]$. After, to get a well-known lower bound on the optimization problem (6), one rounds the optimal solution in a convenient way: the easiest rounding would imply $S_j = 1$ if $S_j^* \geq 0.5$ and $S_j = 0$ if $S_j^* < 0.5$ (and similar for $V_j, \forall j$); also, one could round the relaxed solution in line with optimal quantities, meaning that the product j is included in the assortment if its relaxed order quantity is greater than the minimal lot size.

Both variants would give lower bounds on the optimization problem (6) by evaluation of the objective function with these assortment vectors.

4.4.1. Rank and optimize with dependencies

In order to improve the lower bounds, we propose two heuristics for our bilevel product assortment optimization problem.

First of all, we propose the procedure described in Algorithm 1 to compute tangent points Q_j^s and Q_j^v based on implicit functions (12) using optimal quantizers.

Algorithm 1 Tangent points estimates.

```
Obtain optimal quantizers $\widetilde{\boldsymbol{z}}_j^s = (\widetilde{z}_{1j}^s, \dots, \widetilde{z}_{Nj}^s)$, $\widetilde{\boldsymbol{z}}_j^v = (\widetilde{z}_{1j}^v, \dots, \widetilde{z}_{Nj}^v)$
and corresponding probabilities
$\widetilde{\boldsymbol{p}}_j^s = (\widetilde{p}_{1j}^s, \dots, \widetilde{p}_{Nj}^s)$, $\widetilde{\boldsymbol{p}}_j^v = (\widetilde{p}_{1j}^v, \dots, \widetilde{p}_{Nj}^v)$ for densities $g_j^s(y)$ and $g_j^v(y)$;
for $k = 1, \dots, N$ do
    Estimate the integrals $\int_{Q_j^s}^{\infty} y g_j^s(y) dy$ and $\int_{Q_j^v}^{\infty} y g_j^v(y) dy$ by the
    dot products $(\widetilde{z}_{kj}^s, \dots, \widetilde{z}_{Nj}^s) \cdot (\widetilde{p}_{kj}^s, \dots, \widetilde{p}_{Nj}^s)$ and $(\widetilde{z}_{kj}^v, \dots, \widetilde{z}_{Nj}^v) \cdot (\widetilde{p}_{kj}^v, \dots, \widetilde{p}_{Nj}^v)$;
end for
Further, for the estimates of tangent $Q_j^s$ and $Q_j^v$, choose $\widetilde{z}_{kj}^s$ and $\widetilde{z}_{kj}^v$ with minimal distances
$\min_k \left\{ (\widetilde{z}_{kj}^s, \dots, \widetilde{z}_{Nj}^s) \cdot (\widetilde{p}_{kj}^s, \dots, \widetilde{p}_{Nj}^s) - \frac{(m_j^s + h_j^s)\mu_j^s T_j^s - K_j^s}{cu_j^s} \right\}, \quad \forall j \in \mathcal{N}_s,$
$\min_k \left\{ (\widetilde{z}_{kj}^v, \dots, \widetilde{z}_{Nj}^v) \cdot (\widetilde{p}_{kj}^v, \dots, \widetilde{p}_{Nj}^v) - \frac{(m_j^v + c_j^v + co_j^v)\mu_j^v T_j^v - K_j^v}{cu_j^v + co_j^v} \right\}, \quad \forall j \in \mathcal{N}_v.$
Differently, one could choose the middle point between two closest (in the sense of minimal distance) quantizers.
```

Clearly, as the number of optimal quantizers increases, the accuracy of the approximation improves.

Next, similar to the models considered in Katsifou et al. (2014) and Katsifou (2013), the demand and consequently the profit from a product depend on the selected total assortment in our bilevel model. Therefore, in the solution procedure, the expected profit to space ratio for each selected and non-selected product changes as the total assortment is built. In the extreme case, as a new product is introduced, a previously included product j might become less profitable than other, non-included, products. In another case, the cross-selling opportunity could lead to the introduction of a less profitable but very attractive product to increase overall store traffic. This shows that considering products separate from each other and sequentially selecting them based on their profit to space ratios is, in general, suboptimal, due to the interdependence of demands.

Heuristics proposed in Katsifou et al. (2014) and Katsifou (2013) are based on the *Rank and Approximate* method for the problem without constraints on minimal order quantities and with Monte-Carlo sampling of demand. We, however, propose to adapt the *Rank and Optimize* method using optimal quantizers, taking into account interdependencies between demands. We select a new product depending on the expected profit to space ratio of the current *assortment* and not just on the product's own profit to space ratio. Due to the reoptimization step, we explicitly take minimal order quantities into account.

The heuristic based on the *Rank and Optimize* method is presented in Algorithm 2. The heuristic takes interdependencies between demands into account. First, it ranks the products based on their expected profit to space ratios. Second, it progressively includes products into the assortment choosing the assortment with the maximal profit. In comparison to the heuristic presented in Katsifou et al. (2014) and Katsifou (2013), the Algorithm 2 assigns optimal order quantities to the selected assortment at each iteration. The number of iterations is bounded by $(N_s + N_v + 1)\left\lceil \frac{N_s+N_v}{2} \right\rceil + N_s + N_v$.

Algorithm 2 Rank and Optimize with dependencies

```
Ranking step:

Ranking = {}, Products under Consideration = $\{\mathcal{N}_s \cup \mathcal{N}_v\}$
while numel(Products under Consideration) > 0 do
  for all $j \in$ Products under Consideration do
    $A_j$ = Ranking $\cup\ j$
      • $\forall k \in A_j$, update the demands $\mu_k(A_j)$ for all standard and
        variable products in $A_j$ by solving the problem (3) (note
        that demands are changing depending on the set $A_j$);
      • Using Algorithm 1, compute the maximal profit to space
        ratio of the product $j$, i.e., $r_j = \max_{Q_j}\left(\frac{\Pi_j(Q_j,\mu_j(A_j))}{l_j Q_j}\right)$,
        where $\Pi_j(Q_j, \mu_j(A_j))$ is the profit of the product $j$ depending
        on the set of products $A_j$ through the mean
        function and $Q_j$ is its quantity;
  end for
  Ranking= Ranking $\cup \{j = \arg\max_j r_j\}$;
  Products under Consideration = Products under Consideration \
  $\{j = \arg\max_j r_j\}$.
end while

Reoptimization step:

Current Assortment = {}
for all $j \in$ Ranking do
  $A_j$ = Current Assortment $\cup\ j$
    • Solve the optimization problem (8) given the Current Assortment
      and compute the optimal value $(\Pi_j(Q_j))^*$;
end for
Optimal assortment is $A_j$ : $j = \arg\max_j\{(\Pi_j(Q_j))^*\}$.
```

We use the example from Katsifou et al. (2014) to describe the heuristic in more detail. Let us assume that we have three products $N1$, $N2$ and $N3$. First, we compute the profit to space ratios of all $N_s + N_v$ products considering them in isolation from each other (i.e., same as in *Rank and Optimize* and in *Rank and Approximate* methods). We select product $N3$, which has the highest ratio. The next product included is not necessarily the product with the second highest profit to space ratio, but the one which maximizes the profit to space ratio of the *assortment*. When considering whether to include product $N1$, one updates the demand for and the profit of products $N3$ and $N1$ as a function of the assortment $\{N3, N1\}$. Similarly, when considering $N2$, one updates the demand for and the profit of products $N3$ and $N2$ based on $\{N3, N2\}$. If combination $\{N3, N1\}$ yields a higher profit to space ratio than combination $\{N3, N2\}$, one chooses the product $N1$. Therefore, it may happen that less profitable but more attractive product $N1$ is preferred to $N2$ since its combination with product $N3$ maximizes total profit per space ratio.

In Katsifou et al. (2014), each customer is assumed to buy only one item of a product and neither the optimal purchase quantities of customers nor their budget constraints are taken into account. Via this simplified model for customer purchases, article (Katsifou et al., 2014) presents a profit ratio heuristic for the product assortment optimization with MNL customer choice model and with prices being the decision variables in the retailer's optimization. In our bilevel model, the prices are considered to be fixed, influencing both the customer choice (3) and the retailer's profit (6). Considering prices as decision variables would significantly complicate the problem: the use of the heuristic similar to the one proposed in Katsifou et al. (2014) would not be possible, as the expected demand would tend to infinity with prices being close to zero and the computation of optimal prices based on the expected demand times its profit margin would not be feasible.

Instead of using Algorithm 1 in the ranking step of Algorithm 2, one can use the approximation $\frac{(\Pi_j)^*}{l_j T_j (Q_j)^*}$, where $(\Pi_j)^*$ and $(Q_j)^*$ are the optimal profit and the optimal quantity corresponding to the (standard or variable) product j in the problem (8) with fixed assortment vectors. Further, this method allows to cut the number of iterations by $N_s + N_v$, keeping track of the optimal values already at the ranking step and, after, choosing the maximal one without an additional reoptimization step. In the numerical part, we refer to the Algorithm 2 as *TgRankOptDepend* and to the alternative algorithm with reduced number of iterations and approximated profit to space ratios as *CutEnum*.

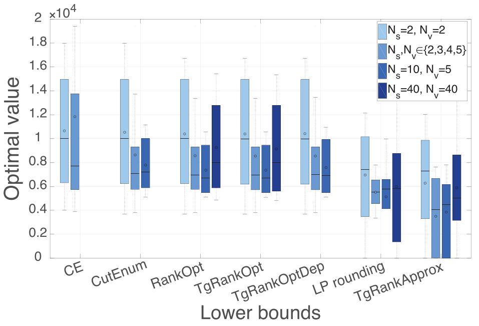
4.5. Accuracy and efficiency analysis

In this section we assess the efficiency and accuracy of lower and upper bounds proposed for the problem (8). We compare the resulting optimal values and the computational times of these algorithms with those of complete enumeration (CE).

We consider small- to large-dimensional problems starting with random selection of sizes from $N_s = \{2, 3, 4, 5\}$, $N_v = \{2, 3, 4, 5\}$ and proceeding to $N_s = 40$, $N_v = 40$ products. Note that CE would need $2^{N_s+N_v} - 1$ iterations each of increasing complexity, which results in more than 1 mln. iterations for $N_s + N_v \geq 20$. Therefore, we use CE for the solution of small-dimensional cases only.

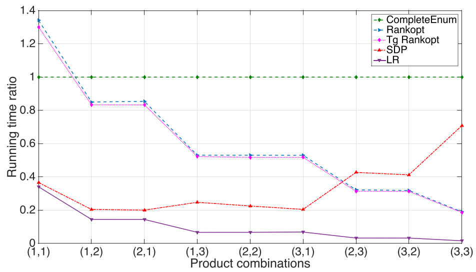
The number of loyal and non-loyal potential customers is $L = \bar{L} = 10$ at every time increment and the store capacity is set at $C = 1500$. All distributions are discretized optimally with $N = 10$ number of quantizers (note that one does not require high number of quantizers for univariate distributions, as the discretization is minimizing the Kantorovich-Wasserschtein distance).

For accuracy and efficiency tests we use parameters listed in Tables 2 and 3. In the customer model, the location and the scale parameters of the budget distribution have a higher spread for non-loyal customers (due to larger uncertainty about them). The risk aversion towards standard products is higher for non-loyal customers, while the risk aversion towards variable products is larger for the loyal group of customers. In the retailer model, the variable products are supposed to offer a lower margin than standard products (40 or 80% of the minimum between the smallest standard products' profit margin and the order cost of the corresponding variable product), and their setup cost is supposed to be higher. By this, variable products tend to be more costly (Table 3) but are mainly intended to attract non-loyal customers to the store. We generate values of the parameters randomly and compute lower and upper bounds for each random parameter instance.



$N_s, N_v \in \{2,3,4,5\}$		
Ratio	**Mean**	**St.dev.**
$\frac{\Pi_{\text{RankOpt}}}{\Pi_{\text{CE}}}$	0.9189	0.1833
$\frac{\Pi_{\text{TgRankOpt}}}{\Pi_{\text{CE}}}$	0.9160	0.1823
$\frac{\Pi_{\text{CutEnum}}}{\Pi_{\text{CE}}}$	0.9322	0.1829
$\frac{\Pi_{\text{TgRankApprox}}}{\Pi_{\text{CE}}}$	0.4536	0.3890
$\frac{\Pi_{\text{TgRankOptDepend}}}{\Pi_{\text{CE}}}$	0.9196	0.1811
$\frac{\Pi_{\text{LRrounding}}}{\Pi_{\text{CE}}}$	0.6399	0.3312

Fig. 2. Optimal values and approximation error for problems with different sizes.



$N_s, N_v \in \{2,3,4,5\}$		
Ratio	**Mean**	**St.dev.**
$\frac{T_{\text{RankOpt}}}{T_{\text{CE}}}$	0.1503	0.1230
$\frac{T_{\text{TgRankOpt}}}{T_{\text{CE}}}$	0.1484	0.1214
$\frac{T_{\text{CutEnum}}}{T_{\text{CE}}}$	0.2605	0.1556
$\frac{T_{\text{TgRankApprox}}}{T_{\text{CE}}}$	0.0832	0.0742
$\frac{T_{\text{TgRankOptDepend}}}{T_{\text{CE}}}$	0.3294	0.2119
$\frac{T_{\text{LRrounding}}}{T_{\text{CE}}}$	0.0358	0.0405

Fig. 3. Ratio of running times for problems with different sizes.

Table 2
Customer model parameter values.

Parameter	Loyal customers	Non-loyal customers
μ, $\overline{\mu}$	$U[7.5, 8.5]$	$U[6.5, 8.5]$
σ, $\overline{\sigma}$	$U[0.45, 0.55]$	$U[0.4, 0.6]$
γ, $\overline{\gamma}$	$U[0.05, 0.07]$	$U[0.6, 0.7]$
δ, $\overline{\delta}$	$U[0.9, 0.95]$	$U[0.5, 0.6]$

Table 3
Retailer model parameter values.

Parameter	Standard products	Variable products
$Q^s_{j,\min}$, $Q^v_{j,\min}$	1	1
l^s_j, l^v_j	$U[0.1, 0.5]$	$U[0.1, 1]$
c^s_j, c^v_j	$l^s_j U[50, 100]$	$l^v_j U[50, 100]$
m^s_j, m^v_j	$\{20\%, 30\%, 40\%\}$ of c^s_j	$\{40\%, 80\%\}$ of $\min\left\{\min_{i=1,\ldots,N_s}\{m^s_i\}, c^v_j\right\}$
K^s_j, K^v_j	500% of c^s_j	750% of c^v_j
u^s_j, u^v_j	50% of m^s_j	30% of m^v_j
cu^s_j, cu^v_j	$m^s_j + h^s_j + u^s_j$	$m^v_j + c^v_j + u^v_j$
co^v_j	–	$\{5\%, 10\%, 25\%\}$ of c^v_j
h^s_j	c^s_j	–
T^s_j, T^v_j	$\frac{2'000}{\hat{\mu}^s_j l^s_j}$	$\frac{2'000}{\hat{\mu}^v_j l^v_j}$

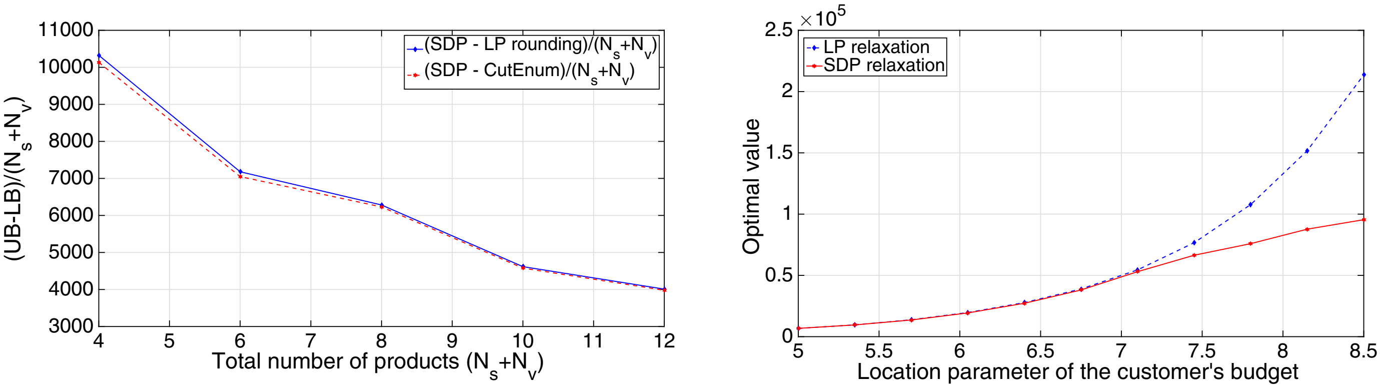
In Figs. 2 and 3 (and from the corresponding tables), we see that lower bounds *RankOpt* and *TgRankOpt* with optimal quantizers are able to find solutions within 10% accuracy in reasonable time compared to CE (both of the bounds are linear in the number of iterations w.r.t. the problem dimension). This ability to find good solutions in short time is important when larger (realistic) size instances have to be solved.

The heuristics *CutEnum* and *TgRankOptDepend*, taking interdependence between assortment and the demand explicitly into account at each iteration, are, obviously, slower than *RankOpt* and *TgRankOpt* (they have quadratic number of iterations w.r.t. the problem dimension): these bounds, however, improve the accuracy of the solution. Further, the lower bound *TgRankApprox*, based on the algorithm developed in Katsifou et al. (2014), provides a very fast solution estimate, which, however, drops below 50% accuracy with high standard deviation for our bilevel problem with minimal lot sizes. Note that simply rounding the LP relaxation, one would get a more accurate estimate in even better time. For higher dimensional cases, as one can see in Fig. 2 for $N_s = N_v = 40$, the performance of both bounds *TgRankApprox* and of *LP rounding* improves, while the running time stays much lower for the *LP rounding*, making this method very time-efficient even with optimal quantization.

Comparing the performance of two upper bounds, one can immediately observe that the LP relaxation becomes more time-efficient if the dimension of the problem increases, while the time-efficiency of the SDP relaxation drops compared to CE (see Fig. 3): this is due to the fact that the number of variables in the SDP relaxation grows quadratically with an increase in dimension (i.e., the number of elements of matrix $\mathbf{X}$ increases). To improve the efficiency of the SDP relaxation, one would require separate research on faster computational algorithms for its solution.

Next, Fig. 4(a) demonstrates the behavior of the upper (SDP) bound w.r.t. lower bounds (*LP with rounding* and *CutEnum*) and w.r.t. the number of products. Fig. 4(b) shows the optimal values of the upper bounds w.r.t. the location parameter of the customer's budget distribution: the performance of the LP relaxation drops w.r.t. the performance of the SDP relaxation as the budget of customers increases (Fig. 4(b)).

In order to further improve the upper bounds for large-scale problems, one would need to find a way to estimate the mean demand close to its optimum in the initial problem instead of taking ranges $0 \leq \mu^s_j \leq \hat{\mu}^s_j$, $0 \leq \mu^v_j \leq \hat{\mu}^v_j$, $\forall j$ in the relaxations. Also, one could impose even stronger constraints (rather than LP constraints introduced by Fortet (1960) or Sherali-Adams constraints (Sherali & Adams, 1990) for the SDP relaxation) to keep assortment vectors as close to a binary solution as possible.



a) Convergence

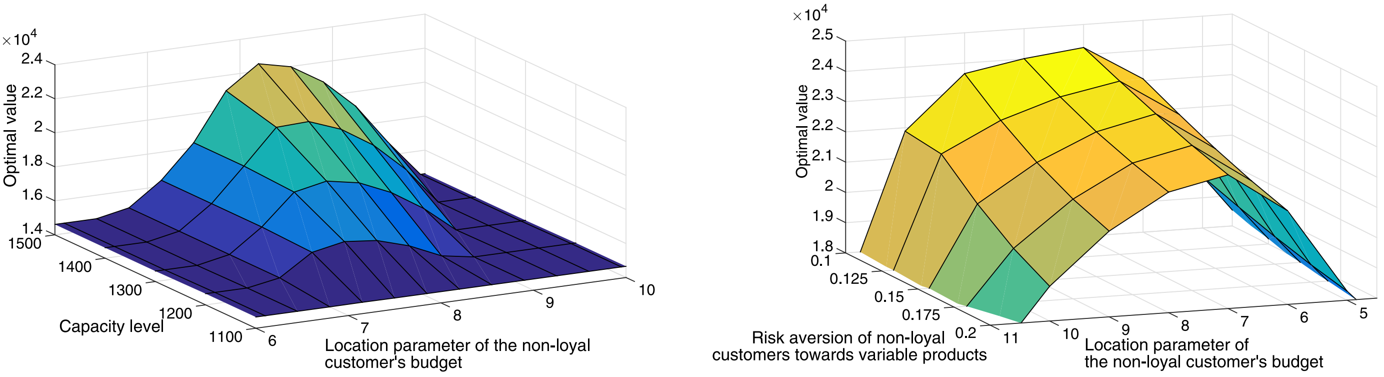
b) Optimal values

Fig. 4. Behavior of the bounds dependent on the number of products and on the location parameter of the customer's budget distribution.

4.6. Functional properties

One of the main contributions of our work is the ability of the model to account for optimal purchase quantities of customers under their budget constraints. Moreover, the budget of a particular customer is unknown to the retailer, who estimates the budget distribution function. Increasing the location parameter of the distribution function implies that the average customer is able to purchase more quantities and a broader set of products. While most customers can be satisfied, this leads to an increase in the retailer's profit. However, if the budget of the average customer increases too much and the capacity of the retailer is not adjusted, the retailer would start facing higher loss-of-goodwill costs and would observe a decrease in his profits: this is due to the retailer's inability to satisfy customers with very high demands (Fig. 5). In order to satisfy more customers, the retailer would need to increase his capacity level (Fig. 5a)). If the customers from the same budget group would become more risk-averse towards purchasing the products, the retailer's profit would logically decrease (Fig. 5b)).

Next, computing the profit of a standard or a variable product j requires the review period (T_j^s or T_j^v). Introducing the review periods as the decision variables would make the optimization problem not convex-concave. Furthermore, for marketing reasons, the review periods are often decided beforehand. We use $T_j^s = \frac{2k^s}{l_j^s \hat{\mu}_j^s}$, $\forall j \in \mathcal{N}_s$ and $T_j^v = \frac{2k^v}{l_j^v \hat{\mu}_j^v}$, $\forall j \in \mathcal{N}_v$ as the approximations



a) Higher customer demands.

b) Customers' risk-aversion.

Fig. 5. Optimal purchase quantities for different types of customers.

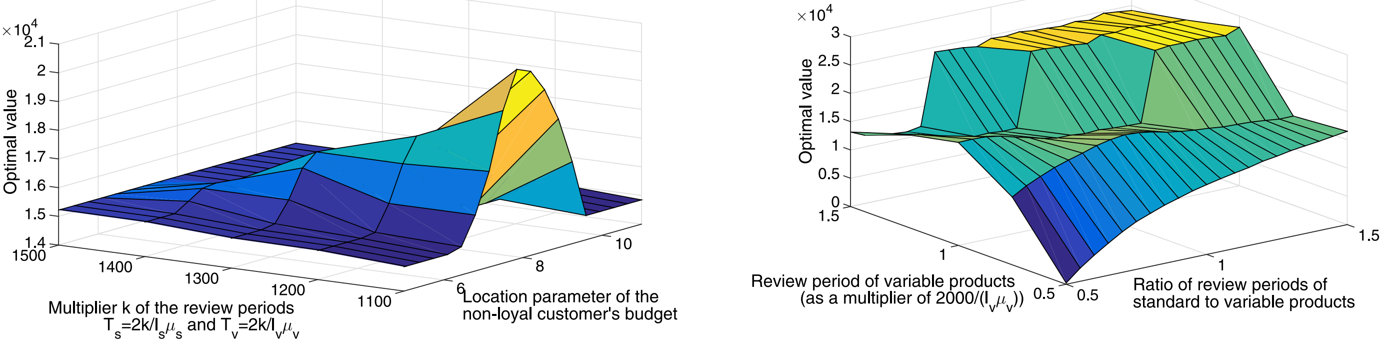


Fig. 6. Optimal purchase quantities for different review periods.

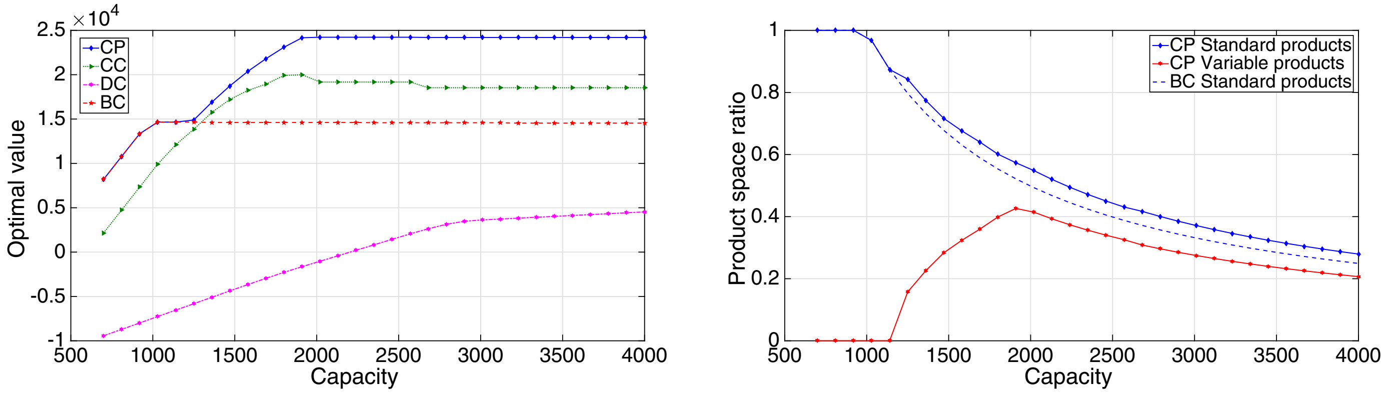


Fig. 7. Profit of scenarios and the proportion of space for standard and variable products when the shelf space increases (less cross-selling).

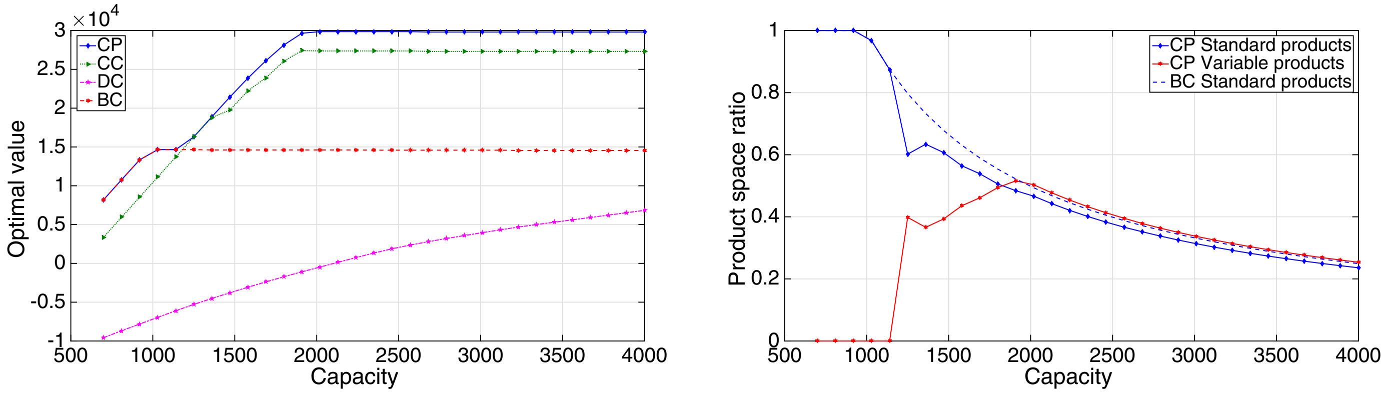


Fig. 8. Profit scenarios and the proportion of space for standard and variable products when the shelf space increases (more cross-selling: $\delta = 0.7$, $\bar{\delta} = 0.2$).

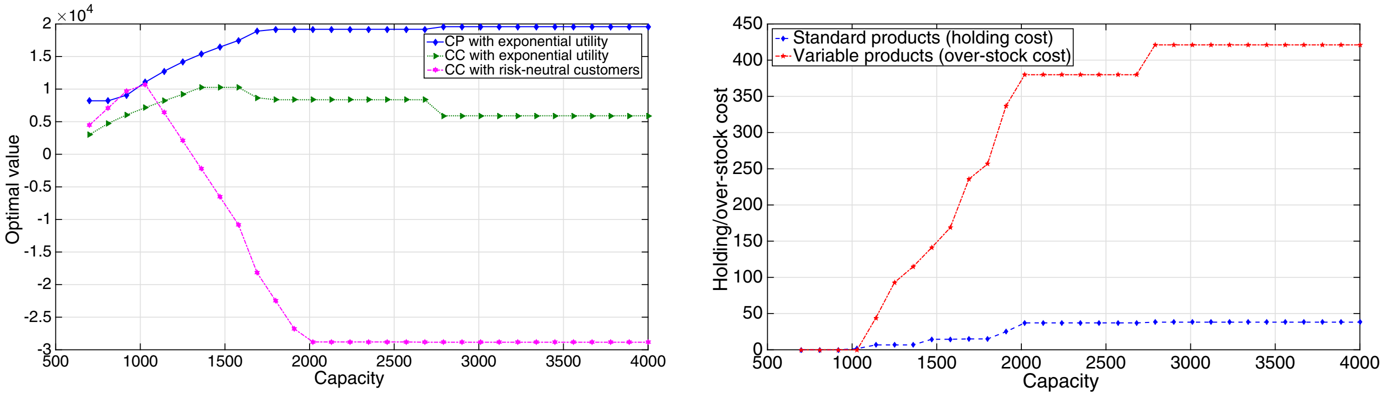


Fig. 9. Profit of CP and CC scenarios when the shelf space increases.

of the review periods in our numerical tests with k^s and k^v being some constants. We choose the parameters k^s and k^v in such a way that the retailer's overall profit is as high as possible. As expected, in Fig. 6 we observe that the problem is not convex-concave in review periods with a maximum dependent on the customers' budget. While it was not our objective to maximize the profit based on the ratio of review periods, the observation that variable products should be refreshed more often than standard ones ($T_j^v < T_j^s$) coincides with examples of retailers in practice (Appendix 1.4).

5. Benchmark strategies

In this section, our aim is to illustrate how our model can contribute to retail practice, by providing finely constructed assortments to attract loyal and non-loyal customers, and to balance profitable and attractive products. For this, we compare our approach with benchmark strategies commonly used in practice, and easier to solve due to their simplified structure. The scenarios are presented in the following.

Base case (BC): The store carries only a standard assortment and thus only attracts loyal customers. Variable products are not considered by the store manager.

Crude case (CC): Both standard and variable items are carried and cross-selling occurs in the store, affecting the demand for each product. However, for simplicity, when product assortment and inventory decisions are taken, the cross-selling effect is ignored, i.e., the assortments are decided based on the assumption that loyal (respectively non-loyal) customers buy only standard (variable) products.

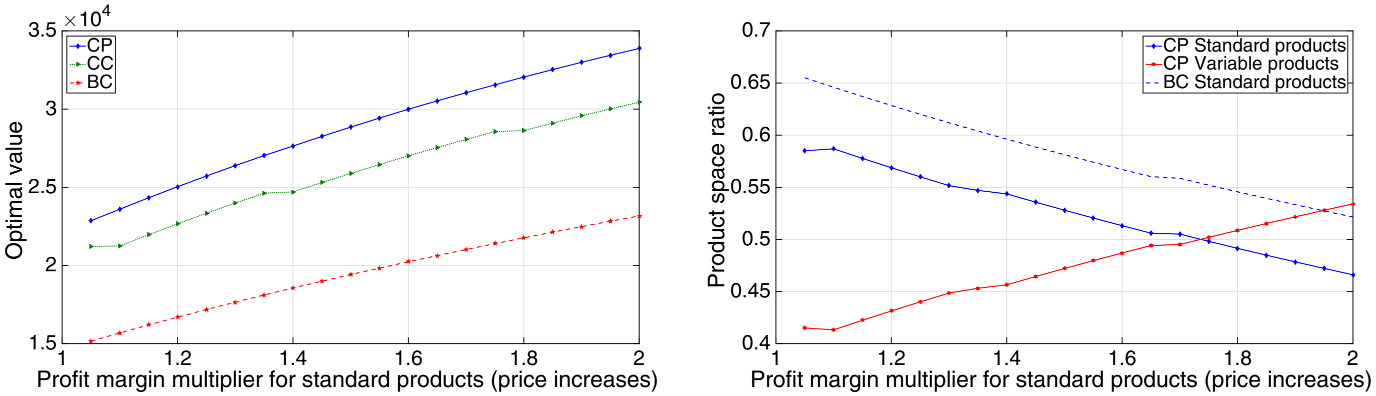


Fig. 10. Profit and the product space ratio when the standard products' price increases.

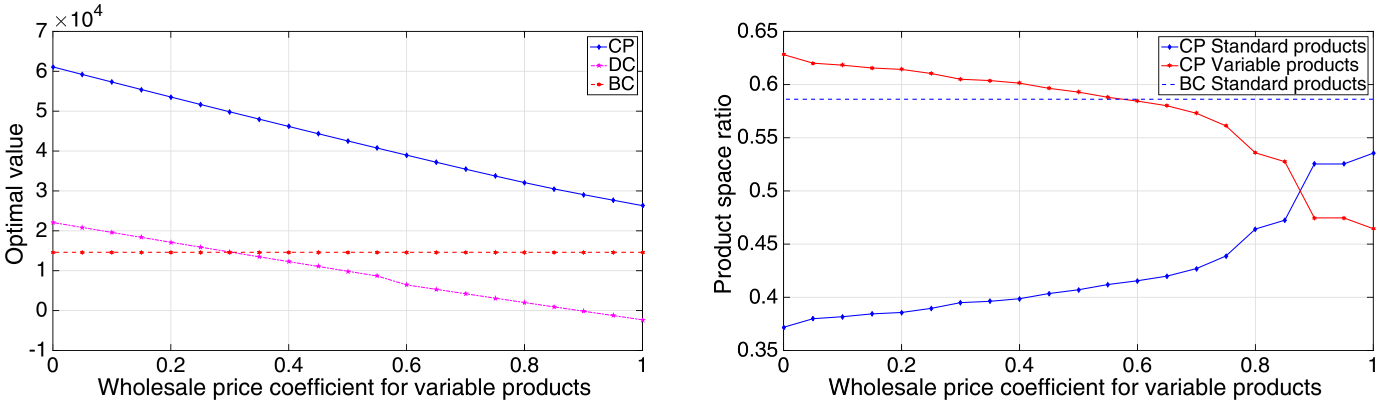


Fig. 11. Profit of scenarios and the proportion of space for standard and variable products when the wholesale price coefficient for variable products changes.

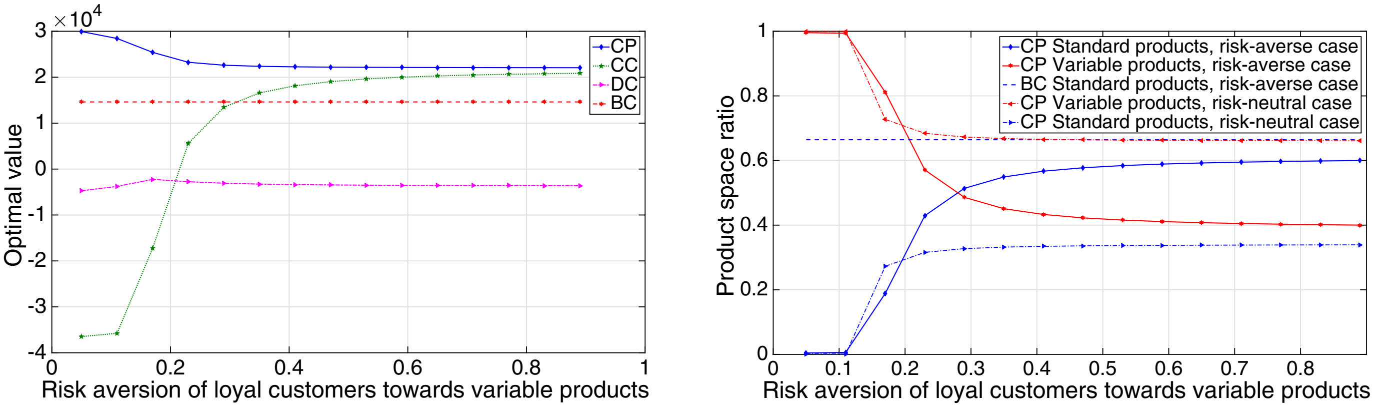


Fig. 12. Profit of scenarios and the space proportion of products when the risk aversion of loyal customers towards variable poducts increases (capacity is fixed at $C = 1500$).

Dogmatic case (DC): The maximal space allocated to each of the standard and the variable products is fixed and limited a priori, from an external decision in the following way:

$$Q_j^s l_j^s \leq S_j \frac{C\hat{\mu}_j^s}{\sum_{j=1}^{N_s} \hat{\mu}_j^s + \sum_{j=1}^{N_v} \hat{\mu}_j^v}, \quad j \in \mathcal{N}_s \quad \text{and}$$

$$Q_j^v l_j^v \leq V_j \frac{C\hat{\mu}_j^v}{\sum_{j=1}^{N_s} \hat{\mu}_j^s + \sum_{j=1}^{N_v} \hat{\mu}_j^v}, \quad j \in \mathcal{N}_v.$$

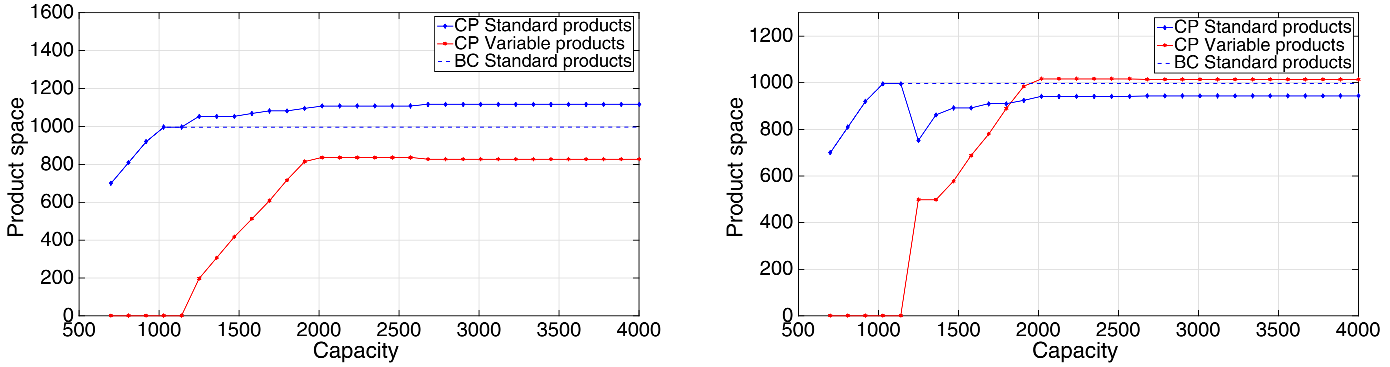
The standard and the variable products are to be placed in this space and their inventory levels are decided based on the customers they attract.

Complete problem (CP): This scenario corresponds to our complete model. The combined assortment and the inventory level for each product are decided taking into consideration the cross-selling effect. The space allocation between the standard and the variable assortments is not fixed a priori.

In the following, we show the benefits brought by our approach for a balanced assortment, and analyze the impact of the space constraint, profit margin and cross-selling effect on the results of the aforementioned scenarios. For the tests, we assume that there are 40 standard and 40 variable candidate products.

5.1. Impact of the space constraint

We demonstrate the results of our approach (CP), and also compare the results of the BC, CC, DC and CP when the shelf space constraint varies in the interval [700, 4000]. Fig. 7 shows that,



a) Risk-aversion $\delta = 0.9$ and $\overline{\delta} = 0.5$, higher order cost for variable products

b) Risk-aversion $\delta = 0.7$ and $\overline{\delta} = 0.2$, lower order cost for variable products

Fig. 13. Space allocated to standard and variable products w.r.t. capacity.

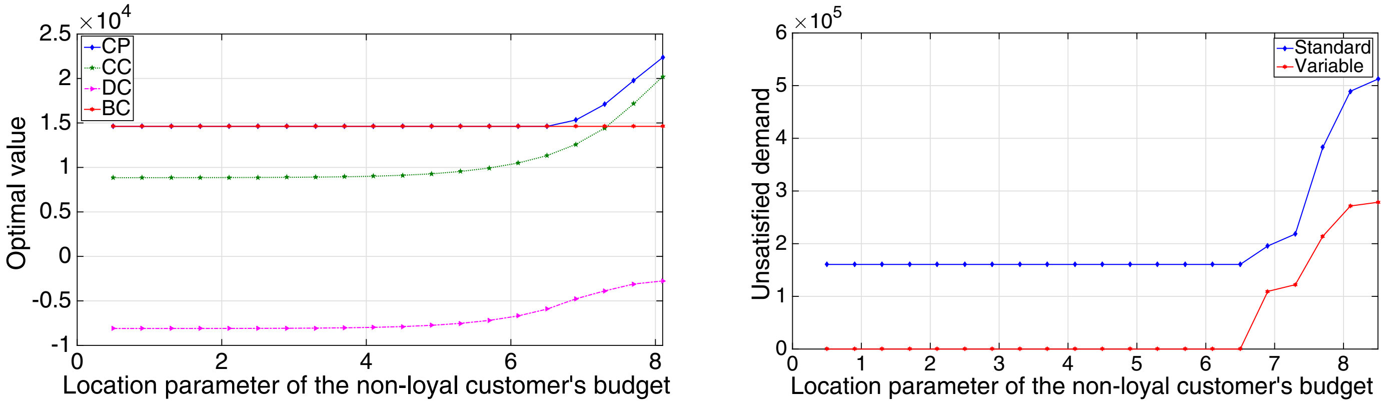
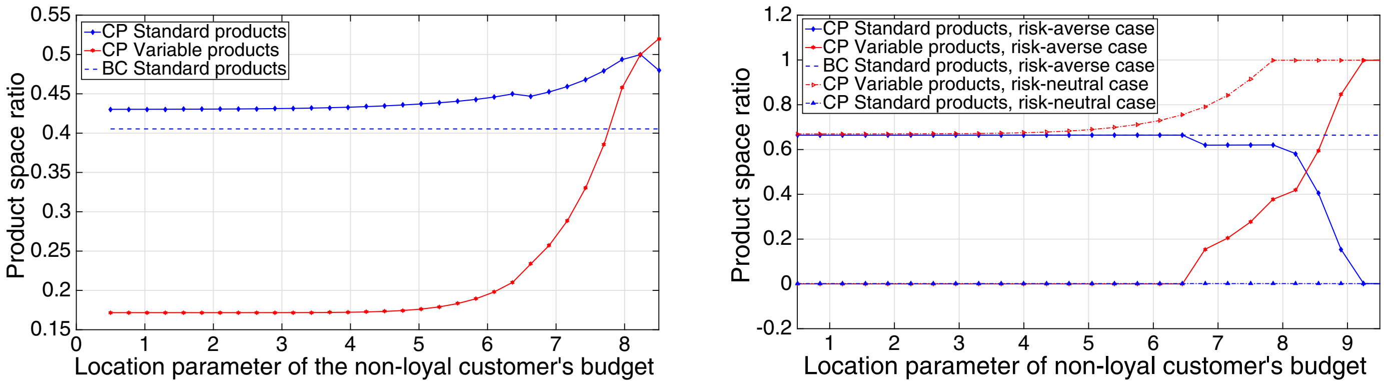


Fig. 14. Profit of scenarios and the unsatisfied demand for standard and variable products when the location parameter of the non-loyal customer's budget increases.



a) Exponential utility.

b) Isoelastic utility.

Fig. 15. The proportion of space for standard and variable products when the location parameter of the non-loyal customer's budget increases for different utility models.

as expected, CP outperforms all the benchmarks in all instances, and in particular the base case (BC) where no variable items are included. Thus, the retailer benefits by including variable products to his assortment and considering the increased demand for standard products due to variable products (cross-selling effect) when deciding on the combined assortment. However, if the capacity is low, the optimal profit of our approach CP may coincide with the optimal profit of the base case BC, meaning that small shops may not necessarily benefit from including variable products. Note that we do not consider the case of not including any of the products and, therefore, optimal values may in general become negative.

In Fig. 7(a), the profit difference between the CP and the BC increases until some capacity level is reached as more variable products can be included in the assortment, increasing the cross-selling effect. Further, the profit difference between CP and BC converges to some level due to the concavity of our optimization problem. The shelf space at which the retailer stops filling the shop up to full capacity is lower for the BC case, due to the absence of the cross-selling effect and impossibility of benefiting further from an increase in order quantities for standard products. Also, due to the constraints on minimal order quantities $Q^s_{j,\min}$, $\forall j = 1, \dots, N_s$ and $Q^v_{j,\min}$, $\forall j = 1, \dots, N_v$, the decision to include variable products to the assortment in CP may delay with respect to the point at which the retailer stops filling the shop up to full capacity in the base case BC.

The impact of the space constraint on the structure of the combined assortment, and the increasing number of variable items included, can also be seen in Fig. 8, where we reduce the wholesale price for variable products and increase the cross-selling effect by setting the risk-aversion parameters towards variable products to $\delta = 0.7$ and $\overline{\delta} = 0.2$. With this, the space ratio devoted to variable products becomes higher than the space ratio devoted to standard products at some capacity level (Fig. 8).

Next, we observe that the crude case CC outperforms the dogmatic approach DC. This is due to the fact that DC is limited to the dedicated space and CC is not. Importantly, the profit of the retailer adopting the crude approach for his assortment decision can decrease with increasing capacity due to the growing holding/overstock cost resulting from the neglection of the cross-selling effect (Fig. 9). In general, the performance of CC and DC strongly depends on the strength of the cross-selling effect.

5.2. Impact of profit margin

To gain additional insight into the combined problem of product assortment and inventory management, we compare the performance of our approach and of the benchmarks when profit margin varies. The standard and variable products' profit margins $p^s_j - c^s_j$ and $p^v_j - c^v_j$ may change due to two reasons: either the price of the product changes or the ordering cost starts to vary. For standard assortment, both of those lead to a change in the profit margin $m^s_j = p^s_j - c^s_j$ and in the parameter $cu^s_j = m^s_j + h^s_j + u^s_j$ corresponding to the loss-of-goodwill. Furthermore, the change in price (not in ordering cost though) influences the mean demand through the customers' choice model. For variable assortment, the situation changes: price changes still influence the mean demand through the customers' choice model; however, changes in the order cost c^v_j do not influence the parameter $cu^v_j = p^v_j + u^v_j$ corresponding to the loss-of-goodwill. Thus, one needs to consider at least two cases in order to study the impact of changes in the profit margin: the case, when the product price varies (Fig. 10) (Casado & Ferrer, 2013) and the case, when the ordering cost changes (Fig. 11).

The space constraint is fixed at $C = 1500$. Figs. 10 and 11 confirm that our approach (CP) outperforms the BC. Under the CP scenario, the order quantities of standard products are reduced due to the decrease in purchase ability of customers and due to the drop in demand if the standard products' profit margins are high due to the price increase (Casado & Ferrer, 2013). In response, the quantities of variable products increase and thus make the majority of the assortment (Fig. 10).

Next, we start to vary the wholesale price (order cost) for variable products instead. Changes in the order cost c^v_j do not influence the parameter $cu^v_j = p^v_j + u^v_j$ corresponding to the loss-of-goodwill and, moreover, they do not influence the customers' choice model and the mean demand. A decrease in the wholesale price for variable products makes it beneficial for the retailer to include more in number and in order quantity of variable products compared to the case with the price increase (Fig. 11).

As seen in Figs. 10 and 11, the scenarios (CC and DC) are outperformed by our approach CP. The performance of the dogmatic case DC improves if the wholesale price for variable products decreases. For lower values of the order cost for variable products, the DC outperforms the base case BC: this is due to the fact that variable products become much more profitable, so that the space limited by the dogmatic approach still leads to a better payoff than the payoff of not including variable products at all.

5.3. Impact of the cross-selling effect

The cross-selling effect is one of the most important characteristics of our problem. We analyze the impact of loyal customers' risk aversion parameter δ towards variable products on optimal profits for CP, BC, CC and DC and on optimal product space ratios. The lower the risk aversion towards variable products, the more customers will buy variable products: varying δ one influences the number of loyal customers buying variable products and the quantities they purchase.

First, Fig. 12 confirms that our approach (CP) consistently outperforms the BC. With our model, the retailer benefits by considering the cross-selling effect when he is deciding on the total assortment and the inventory level per product. As the loyal customers become more willing to buy variable products, the benefit that the retailer gains with our approach increases. One can see that as the risk aversion parameter δ drops, the cross-selling effect increases and the retailer increases the quantity of variable products in the assortment in response to the gain in their demand (otherwise, the loss-of-goodwill for variable products would be too high). A similar effect is observed for the case when a part of customers is risk-neutral: in Fig. 12, the retailer places more emphasis on the variable assortment as a response to the risk-neutral customers' purchasing power.

Next, we observe that the scenario CC converges towards our approach (CP) as the risk aversion of loyal customers for variable products increases: this is due to the fact that high risk aversion parameters imply low cross-selling also in the complete problem CP. At the same time, the scenario CC starts outperforming base case BC after some level of risk aversion parameter, meaning that even low cross-selling effect generates positive profit by inclusion of variable assortment. The dogmatic approach DC, limiting space allocated for each of the products, is outperformed by all other methods, being, furthermore, outperformed by the case with no products included (zero payoff).

The impact of the cross-selling effect on the structure of the combined assortment, and the increasing number of variable items included, can also be seen in Fig. 13, where we vary the wholesale price for variable products and reduce the risk-aversion parameters towards variable products for both loyal and non-loyal customers ($\delta = 0.7,\ \overline{\delta} = 0.2$). The space devoted to variable products outperforms the space devoted to standard ones at some capacity level in case of lower risk aversion.

Further parameters having influence on the cross-selling effect are the customers' budget parameters: if the location parameter of the customers' budget distribution is too low, it does not allow to purchase enough quantities of the products, which influences the mean demand. For different utility models, changing the location parameter $\overline{\mu}$ of the lognormal distribution for non-loyal customers with fixed scale parameter $\overline{\sigma} = 0.5,$ we obtain Figs. 14 and 15, clearly suggesting an increase in variable product quantities as the response to the increase in customers' budget. From this also follows that the assortment consisting to a large extent of variable products is optimal in the absence of non-loyal customers' budget constraint and independently of the utility type (note that isoelastic utility would imply variable assortment only).

Clearly, as the location parameter of the non-loyal customer's budget distribution increases, non-loyal customers may purchase more quantities of the products, increasing the demand. Due to the fixed capacity, the unsatisfied demand for both types of products also increases, as seen in Fig. 14. However, we observe that optimal unsatisfied demand ratios are stable after variable products being included to the assortment (before this, all unsatisfied demand corresponds to standard products, as there is no demand for variable products in the base case BC).

6. Conclusion

This paper contributes in different aspects to the joint problem of product assortment planning and inventory management optimization. We consider an assortment consisting of “standard” and “variable” products serving “loyal” and “non-loyal” customers with uncertain budgets. The cross-selling effect may occur for both groups of customers, encouraging different customers to purchase products from different groups. We describe the strength of the cross-selling effect via coefficients of relative risk aversion towards standard and variable products and we distinguish between budget distributions and budget constraints for each group of customers, limiting baskets that different customers are able to purchase. Next, we formulate and solve a bilevel optimization problem under decision-dependent uncertainty, where the retailer’s binary decision about product inclusion influences the distribution of the product’s demand. Each customer has his own optimization problem to solve and the optimal solution depends on the assortment available at the retailer and optimized under the store’s shelf space constraint. The optimal solution of the customer’s optimization problem is the quantity to purchase under the budget constraint. For the numerical solution, we propose heuristic algorithms, providing lower bounds on the optimal value of the optimization problem and taking the interdependency between the assortment and the demand explicitly into account. We make a comparison to existing lower bounds and we formulate two upper bounds for the bilevel optimization problem via LP and SDP relaxations. Proposed heuristics are more accurate than other lower bounds in finding the optimal solution, being quadratic in the number of iterations. Using optimal quantization techniques instead of typical Monte-Carlo simulations for the approximation of demand and customer’s budget distributions, we enhance both accuracy and efficiency of the solution, as the necessary number of quantizers is much lower than in Monte-Carlo sampling. Using the optimal quantization for the lower bounds allows to devise high-dimensional assortments. For very high-dimensional problems, one could round the optimal solution of the LP upper bound, which would result in a very fast solution estimate. According to our results, a retailer benefits from a combined assortment if he selects it taking into consideration the cross-selling effect that occurs. Conversely, if he includes variable products in the assortment and ignores the cross-selling effect, he might lose a significant amount of profit. In this case, it can be even more profitable for him to carry only standard products; this, however, strongly depends on the customer’s budget distribution, as too high optimal demands of customers lead to a decrease in the retailer’s expected profit due to the increase in lost sales. The variety of techniques proposed in our work can be used to address more complete product assortment strategies in the future. Our results can be extended by including price decisions or dynamic product substitution. Further, in the presence of numerical data on demand, the proposed bounds can be tightened via introduction of constraints on mean demands.

Acknowledgments

We are very grateful to the Austrian Science Fund (FWF Der Wissenschaftsfonds): J3674-N26 for funding our research.

Supplementary material

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.ejor.2020.02.019.

References

Agrawal, N., & Smith, S. (1996). Estimating negative binomial demand for retail inventory management with unobservable lost sales. *Naval Research Logistics, 43*, 839–861.

Agrawal, N., & Smith, S. (2003). Optimal retail assortments for substitutable items purchased in sets. *Naval Research Logistics, 50*, 793–822.

Aitchison, J., & Brown, J. A. C. (1957). *The lognormal distribution with special reference to its uses in economics*. Cambridge Univ. Press.

Amit, R., Mehta, P., & Tripathi, R. R. (2015). Optimal shelf-space stocking policy using stochastic dominance under supply-driven demand uncertainty. *European Journal of Operational Research, 246*(1), 339–342.

Anderson, S. P., de Palma, A., & Thisse, J. F. (1992). *Discrete choice theory of product differentiation*. MA: MIT Press.

Bard, J. F. (1991). Some properties of the bilevel programming problem. *Journal of Optimization Theory and Applications, 68*(2), 371–378.

Bard, J. F. (1998). *Practical bilevel optimization: algorithms and applications*. Kluwer Academic Publishers.

Ben-Akiva, M., & Lerman, S. (1994). *Discrete choice analysis: theory and applications to travel demand* (6th Ed.). Cambridge, MA: MIT Press.

Ben-Ayed, O., & Blair, C. E. (1990). Computational difficulties of bilevel linear programming. *Operations Research, 38*(3), 556–560.

Bijvank, M., & Vis, I. F. A. (2011). Lost-sales inventory theory: A review. *European Journal of Operational Research, 215*(1), 1–13.

Billington, C., & Nie, W. (2009). The customer value proposition should drive supply chain design: An example in mass retailing. *Perspectives For Managers, IMD*.

Calvete, H. I., Galé, C., & Mateo, P. M. (2008). A new approach for solving linear bilevel problems using genetic algorithms. *European Journal of Operations Research, 188*(1), 14–28.

Casado, E., & Ferrer, J. C. (2013). Consumer price sensitivity in the retail industry: Latitude of acceptance with heterogeneous demand. *European Journal of Operational Research, 228*(2), 418–426.

Colson, B., Marcotte, P., & Savard, G. (2007). An overview of bilevel programming. *Annals of Operations Research, 153*(1), 235–256.

Davenport, T. H., & Harris, J. G. (2007). *Competing on analytics: the new science of winning*. Harvard Business Press.

Flores, A., Berbeglia, G., & Hentenryck, P. V. (2019). Assortment optimization under the sequential multinomial logit model. *European Journal of Operational Research, 273*(3), 1052–1064. (forthcoming)

Fortet, R. (1960). L’algebre de boole et ses applications en recherche opérationelle. *Trabajos de Estadistica, 11*(2), 111–118.

Gaivoronski, A. A., Lisser, A., Lopez, R., & Xu, H. (2011). Knapsack problem with probability constraints. *Journal of Global Optimization, 49*(3), 397–413.

Garey, M., & Johnson, D. (1979). *Computers and intractability: a guide to the theory of np-completeness*. New York, NY, USA: W. H. Freeman & Co..

Ghoniem, A., & Maddah, B. (2015). Integrated retail decisions with multiple selling periods and customer segments: Optimization and insights, *55*, 38–52.

Grewal, D., Levy, M., Mehrotra, A., & Sharma, A. (1999). Planning merchandising decisions to account for regional and product assortment differences. *Journal of Retailing, 75*(3), 405–424.

Hansen, P., & Heinsbroek, H. (1979). Product selection and space allocation in supermarkets. *European Journal of Operational Research, 3*, 474–484.

Helmberg, C., Rendl, F., & Weismantel, R. (2000). A semidefinite programming approach to the quadratic knapsack problem. *Journal of Combinatorial Optimization, 4*(2), 197–215.

Hochrainer-Stigler, S., Timonina-Farkas, A., Silm, K., & Balkovič, J. (2019). Large scale extreme risk assessment using copulas: An application to drought events under climate change for austria. *Computational Management Science*.

Hübner, A., Kuhn, H., & Kühn, S. (2016). An efficient algorithm for capacitated assortment planning with stochastic demand and substitution. *European Journal of Operational Research, 250*(2), 505–520.

Kantorovich, L. (1942). On the translocation of masses. *C.R. (Doklady) Academy of Sciences URSS (N.S.), 37*, 199–201.

Katsifou, A. (2013). *Variable product portfolio management in retail operations*. Doctoral Dissertation, EPFL.

Katsifou, A., Seifert, R. W., & Tancrez, J. S. (2014). Joint product assortment, inventory and price optimization to attract loyal and non-loyal customers. *Omega, 46*, 36–50. 2014

Kök, G., & Fisher, M. (2007). Demand estimation and assortment optimization under substitution: Methodology and application. *Operations Research, 55*(6), 1001–1021.

Kök, G., Fisher, M., & Vaidyanathan, R. (2009). Assortment planning: Review of literature and industry practice. In N. Agrawal, & S. A. Smith (Eds.), *Retail supply chain management. international series in operations research & management science*. Boston, MA: Springer-Verlag US. 99–154

Kumar, N., & Steenkamp, J. (2007). *Private label strategy: how to meet the store brand challenge*. HBS Press.

Li, Z. (2007). A single-period assortment optimization model. *Production and Operations Management, 16*(3).

Liberatore, M. J., & Luo, W. (2010). The analytics movement: Implications for operations research. *Interfaces, 40*(4), 313–324.

Lim, A., Rodrigues, B., & Zhang, X. (2004). Metaheuristics with local search techniques for retail shelf-space optimization. *Management Science, 50*(1), 117–131.

Mahajan, S., & van Ryzin, G. (2001). Stocking retail assortment under dynamic substitution. *Operations Research, 49*(3), 334–351.

Marquard, W. (2007). *Wal-Smart, what it really takes to profit in a Wal-Mart World*. New York: McGraw-Hill.

Méndez-Diaz, I., Bront, J. M., Vulcano, G., & Zepala, P. (2010). A branch-and-cut algorithm for the latent class logit assortment problem. *Electronic Notes in Discrete Mathematics, 36*, 319–326.

Monge, G. (1781). Mémoire sue la théorie des déblais et de remblais. *Histoire de l'Académie Royale des Sciences de Paris, avec les Mémoires de Mathématique et de Physique pour la même année*, 666–704.

Mou, S., Robb, D. J., & DeHoratius, N. (2018). Retail store operations: Literature review and research directions. *European Journal of Operational Research, 265*(2), 399–422.

Nahmias, S. (2005). *Production and operations analysis*. New York: McGraw-Hill.

Prais, S. J., & Houthakker, H. S. (1971). *The analysis of family budgets*: 4. CUP Archive.

Rabbani, M., Salehi, R., & Farshbaf-Geranmayeh, A. (2017). Integrating assortment selection, pricing and mixed-bundling problems for multiple retail categories under cross-selling. *Uncertain Supply Chain Management, 5*(4).

Rachev, S. T., & Rüschendorf, W. (1998). Mass transportation problems. In *Probability and its applications*. New York: Springer-Verlag, Vol. 1 (Theory). ISBN 0-387-98350-3

Ranyard, J. C., Fildes, R., & Hu, T. I. (2015). Reassessing the scope of OR practice: the influences of problem structuring methods and the analytics movement. *European Journal of Operational Research, 245*(1), 1–13.

Rendl, F., & Sotirov, R. (2003). Bounds for the quadratic assignment problem using the bundle method. Technical report University of Klagenfurt, Universitätsstrasse 65-67, Austria.

Institute, S. (2013). Improving retail decisions with customers analytics. In White paper.

Sherali, H. D., & Adams, W. P. (1990). A hierarchy of relaxations between the continuous and convex hull representations for zero-one programming problems. *SIAM Journal Discrete Mathematics, 3*(3), 411–430.

Sherali, H. D., & Adams, W. P. (1994). A hierarchy of relaxations and convex hull characterizations for mixed-integer zero-one programming problems. *Discrete Applied Mathematics, 52*(1), 83–106.

Smith, S., & Agrawal, N. (2000). Management of multi-item retail inventories systems with demand substitution. *Operations Research, 48*, 50–64.

Smith, S. A. (2009). Optimizing retail assortments for diverse customer preferences. In N. Agrawal, & S. A. Smith (Eds.), *Retail supply chain management. international series in operations research & management science* (pp. 183–206). US, Boston, MA: Springer-Verlag.

Timonina, A. V. (2013). *Multi-stage stochastic optimization: the distance between stochastic scenario processes*. Berlin Heidelberg: Springer-Verlag. 10.1007/s10287-013-0185-3

van Ryzin, G., & Mahajan, S. (1999). On the relationship between inventory costs and variety benefits in retail assortment. *Management Science, 45*, 1496–1509.

Villani, C. (2008). *Optimal transport, old and new*. Springer.

Walters, D., & Hanrahan, J. (2000). *Retail strategy: planning and control*. London: Macmillan.

Yang, M.-H. (2001). An efficient algorithm to allocate shelf space. *European Journal of Operational Research, 131*(1), 107–118.

Zentes, J., Morschett, D., & Schramm-Klein, H. (2007). Strategic retail management: Text and international cases. Gabler.

Further Reading

Graf, S., & Luschgy, H. (2000). Foundations of quantization for probability distributions. *Lecture notes in mathematics, Volume 1730*. Berlin: Springer.

Pflug, G. C., & Pichler, A. (2011). Approximations for probability distributions and stochastic optimization problems. In G. Consigli, M. Dempster, & M. Bertocchi (Eds.), *Springer handbook on stochastic optimization methods in finance and energy* (pp. 343–387). Int. Series in OR and Management Science, Vol. 163(15)

Silver, E., Pyke, D., & Peterson, R. (1998). *Inventory management and production planning and scheduling*. Hoboken: Wiley.