Chapter 8
Assortment Planning: Review of Literature and Industry Practice

A. Gürhan Kök, Marshall L. Fisher, and Ramnath Vaidyanathan

1 Introduction

A retailer's assortment is defined by the set of products carried in each store at each point in time. The goal of assortment planning is to specify an assortment that maximizes sales or gross margin subject to various constraints, such as a limited budget for purchase of products, limited shelf space for displaying products, and a variety of miscellaneous constraints such as a desire to have at least two vendors for each type of product.

Clearly the assortment a retailer carries has an enormous impact on sales and gross margin, and hence assortment planning has received high priority from retailers, consultants and software providers. However, no dominant solution has yet emerged for assortment planning, so assortment planning represents a wonderful opportunity for academia to contribute to enhancing retail practice. Moreover, an academic literature on assortment planning is beginning to emerge. The purpose of this chapter is to review the academic literature on assortment planning, to overview the approaches to assortment planning used by several

This paper is an invited chapter to appear in *Retail Supply Chain Management*, Eds. N. Agrawal and S. A. Smith, Kluwer Publishers.

A.G. Kök
College of Administrative Sciences and Economics, Koç University, Istanbul, Turkey
e-mail: gkok@ku.edu.tr; http://home.ku.edu.tr/gkok/

M.L. Fisher
The Wharton School, University of Pennsylvania, Philadelphia, PA, USA
e-mail: fisher@wharton.upenn.edu

R. Vaidyanathan (✉)
Desautels Faculty of Management, McGill University, Montréal, QC, Canada
e-mail: ramnath.vaidyanathan@mcgill.ca

N. Agrawal, S.A. Smith (eds.), *Retail Supply Chain Management*,
International Series in Operations Research & Management Science 223,
DOI 10.1007/978-1-4899-7562-1_8

retailers so as to provide some examples of practice, and to suggest directions for future research.

Retailers engage in assortment planning because they need to periodically revise their assortment. Several factors require a retailer to change their assortment, including seasons (the fall assortment for an apparel retailer will be different from the spring assortment), the introduction of new products and changes in consumer tastes.

Most retailers segment the stock keeping units (SKU) they carry into groups called categories. For example, for a consumer electronics retailer, a category might be personal computers. Within categories, they will usually define subcategories, such as laptops and desktops within the computer category. (The terminology used varies across retailers e.g. department, class and subclass may be used instead of category and subcategory, but the practice of grouping SKUs with similar attributes for planning purposes is universal.) Retailers focus most of their energy on deciding what fraction of their shelf space and product purchase budget to devote to each category and subcategory. For example, a consumer electronics retailer would worry more about how to divide their resources between laptops and desktops than about which specific models of each to carry, a decision that is usually left to a more junior buyer. The resource allocation decisions are based on their own historical sales in each subcategory, especially whether sales in a subcategory have been trending up or down, together with external information from a variety of sources such as industry shows, vendors and competitor moves.

Given fixed store space and financial resources, assortment planning requires a tradeoff between three elements: how many different categories does the retailer carry (called a retailer's breadth), how many SKUs do they carry in each category (called depth), and how much inventory do they stock of each SKU, which obviously affects their in-stock rate. The breadth vs. depth tradeoff is a fundamental strategic choice faced by all retailers. Some, like department stores, will elect to carry a large number of different categories. Others, such as category killers like Toys 'R Us and Best Buy, will specialize in a smaller number of categories, but have great depth in each category.

We have all had the experience of going into a store looking for a particular product, not finding it, and settling for another similar product instead. This is called substitution, and the willingness of customers to substitute within a particular category is an important parameter in assortment planning. If customers have a high propensity to substitute in a category, then providing great depth and a high in-stock rate is less critical. The reverse is also true.

We can delineate three patterns with respect to customer substitution: (1) the customer shops a store repeatedly for a daily consumable and one day she finds it stocked out so she buys another. This is called stock-out based substitution. (2) a customer identifies a favorite product based on ads or what she has seen in other stores, but when she tries to find it in a particular store, she can't because they don't carry it, so see buys another product. This is called assortment based substitution. (3) the consumer chooses her favorite product from the ones she sees on the shelf in a store when she is shopping and buys it if it has higher utility than her no purchase option.

In this case, there may be other products she would have preferred, (but she didn't see them either because the retailer didn't carry them or because they were stocked out), and in this sense we can say she substituted, although she may not be aware that these other products exist and hence doesn't herself think of her purchase decision as involving substitution. The first two patterns are common with daily consumables like food and the later with consumer durables like apparel or consumer electronics.

Assortment planning is a relatively new but quickly growing field of academic study. The academic approach to the assortment planning problem rests on the formulation of an optimization problem with which to choose the optimal set of products to be carried and the inventory level of each product. Decisions for each product are interdependent because products are linked in considerations such as shelf space availability, substitutability between products, common vendors (brands), joint replenishment policies and so forth. Most of the literature focuses on a single category or subcategory of products at a given point in time. While a retailer might have a different assortment at each store, the academic literature has focused on determining a single assortment for a retailer, which could be viewed as either a common assortment to be carried at all stores or the solution to the assortment planning problem for a single store.

This chapter begins in Sect. 2 by briefly reviewing four streams of literature that assortment planning models build on: product variety and product line design, shelf space allocation, multi-product inventory systems and a consumer's perception of variety.

In Sect. 3, we discuss empirical results on consumer substitution behavior and present three demand models used in assortment planning: the multinomial logit, exogenous demand and locational choice models.

In Sect. 4, we describe optimization based assortment planning studies. Sections 4.1–4.3 review optimization approaches for the basic assortment planning problem. The models and solution methodologies in these papers vary because of differences in the underlying demand model and the application context. We then review variations on the basic assortment planning problem, including assortment planning with supply chain considerations in Sect. 4.4, assortment planning with demand learning and assortment changes during the selling season in Sect. 4.5, and multi-category assortment planning that considers the interactions between different categories due to existence of basket shopping consumers in Sect. 4.7.

In Sect. 5, we discuss demand and substitution estimation methodologies. The methods depend on the demand model and the type of data that is available.

In Sect. 6, we present industry approaches to assortment planning. We describe the assortment planning process at four prominent retailers: Electronics retailer Best Buy, book and music retailer Borders, Indian jewelry retailer Tanishq, and Dutch supermarket chain Albert Heijn. As will be seen, these companies take significantly different approaches and emphasize different aspects of the assortment problem.

In Sect. 7, we provide a critical comparison of the academic and industry approaches and use this to identify research opportunities to bridge the gap between the two approaches.

For an earlier overview of the assortment planning literature, see Mahajan and van Ryzin (1999).

2 Related Literature

In this section, we briefly review the literature on topics related to assortment planning.

2.1 Product Variety and Product Line Design

Product selection and the availability of products has a high impact on the retailer's sales, and as a result gross profits and assortment planning has been the focus of numerous industry studies, mostly concerned with whether assortments were too broad or narrow. Retailers have increased product selection in all merchandise categories for a number of reasons, including heterogeneous customer preferences, consumers seeking variety and competition between brands: Quelch and Kenny (1994) report that the number of products in the market place increased by 16 % per year between 1985 and 1992 while shelf space expanded only by 1.5 % per year during the same period. This has raised questions as to whether rapid growth in variety is excessive. For example, many retailers are adopting an "efficient assortment" strategy, which primarily seeks to find the profit maximizing level of variety by eliminating low-selling products (Kurt Salmon Associates 1993), and "category management," which attempts to maximize profits within a category (AC Nielsen 1998). There is empirical evidence that variety levels have become so excessive that reducing variety does not decrease sales (Dreze et al. 1994; Broniarczyk et al. 1998; Boatwright and Nunes 2001). And from the perspective of operations within the store and across the supply chain, it is clear that variety is costly: a broader assortment implies less demand and inventory per product, which can lead to slow selling inventory, poor product availability, higher handling costs and greater markdown costs.

The literature that studies the economics of product variety is vast. The main model in this field is the oligopoly competition between single product firms based on Hotelling (1929). In the Hotelling model, consumers are distributed uniformly on a line segment and firms choose their positions on the line segment and their prices to maximize profits. Consumers' utility from each firm is decreasing in the firm's price and their physical distance to the firm. Each consumer chooses the firm that provides her the maximum utility. The objective is to find the number of firms, their locations and their prices in equilibrium and the resulting consumer welfare. Extensions of this model are used to study product differentiation. There are two types of product differentiation. In a horizontally differentiated market, products are different in features that can't be ordered. In that case, each of the products is ranked first for some of the consumers. A typical example is shirts of different color. In a vertically differentiated market, products can be ordered according to their objective quality from the highest to the lowest. A higher quality product is more desirable than a lower quality product for any consumer. Anderson et al. (1992) and Lancaster (1990) provide excellent reviews of this literature.

One of the outgrowths of the literature on the economics of product variety is the product line design problem pioneered by Mussa and Rosen (1978) and Moorthy (1984). A monopolist chooses a subset of products from a continuum of vertically differentiated products and their prices to be sold in a market to a variegated set of customer classes in order to maximize total profit. Consider cars as a product with a single attribute, say engine size. The monopolist's problem is to choose what size engines to put in the cars and how to price the final product. These papers assume convex production costs and do not consider operational issues such as fixed costs, changeover costs, and inventories. Joint consideration of marketing and production decisions in product line design is reviewed by Eliashberg and Steinberg (1993). Dobson and Kalish (1993) propose a mathematical programming solution for this problem in the presence of fixed costs for each product included in the assortment. Desai et al. (2001) study the product line design problem with component commonality. Netessine and Taylor (2007) extend Moorthy's (1984) work by using the Economic Order Quantity (EOQ) model to incorporate economies of scale. de Groote (1994) also considers concave production costs and analyzes the product line design problem in a horizontally differentiated market. He shows that the firm chooses a product line to cover the whole market and the product locations are equally spaced. Alptekinoğlu (2004) extends this work to two competing firms, one offering infinite variety through mass customization and the other limited variety under mass production. He shows that the mass producer needs to reduce variety in order to mitigate the price competition. Chen et al. (1998) is the only paper that considers product positioning and pricing with inventory considerations. They show that the optimal solution for this model under stochastic demand can be constructed using dynamic programming.

These models were early treatments of assortment planning from the manufacturer's view that were precursors of similar models developed for retailing. The manufacturer's problem is one of product positioning in an attribute space (quality or some other attribute) and pricing. The retailer's problem is to select products from the product lines of several manufacturers. A more careful consideration of inventories at product level is needed in retail assortment planning, since inventories have a direct impact on both sales and costs for the retailer.

2.2 Multi-Item Inventory Models

Multi-item inventory problems are also highly relevant to the assortment planning problem. The inventory management of multiple products under a single a shelf space or budget constraint is studied extensively in the operations literature and solutions using Lagrangian multipliers is presented in various textbooks, e.g., Hadley and Whitin (1963). Downs et al. (2002) describe a heuristic approximation to the multi-period version of this problem with lost sales. In these models, the demand of products are not dependent on others' inventory levels (i.e., there is no substitution between products).

The other group of inventory models with multiple products consider stock-out based substitution, focusing on the stocking decisions given a selection, but not the selection of the products. These models are based on an exogenous model of demand which we shall describe in the next section. Briefly, the total demand of a product is the sum of its own initial demand and the substitution demand from other products. Substitution demand from product k to j is a fixed proportion α_{kj} of the unsatisfied demand of product j. McGillivray and Silver (1978) first introduced the problem with two products. Parlar and Goyal (1984) study the decentralized version of the problem. Noonan (1995) and Rajaram and Tang (2001) present heuristic algorithms for the solution of the case with n products. Netessine and Rudi (2003) investigate the case with n products under centralized and decentralized management regimes. The complexity of the problem is prohibitive and it is not possible to obtain an explicit solution to the problem. Netessine and Rudi (2003) find that a decentralized regime carries more inventory than the centralized regime because of the competition effects. Mahajan and van Ryzin (2001b) establish similar results under dynamic customer substitution with the multinomial logit choice model. Parlar (1985) and Avsar and Baykal-Gursoy (2002) study the infinite horizon version of this problem under centralized and competitive scenarios respectively. Lippman and McCardle (1997) consider a single period model under decentralized management, where aggregate demand is a random variable and demand for each firm is a result of different rules of initial allocation and reallocation of excess demand. Bassok et al. (1999) consider an alternative substitution model, in which the retailer observes the entire demand before allocating the inventory to products. In this retailer controlled substitution model, the retailer may upgrade a customer to a higher quality product. The reallocation solution is obtained by solving a transportation problem.

The literature on assemble-to-order systems is also related. The demand for individual components are linked through the demand for finished goods. See Song and Zipkin (2003) for a review. An online retailer's order fulfillment problem when customers can order multiple products can be viewed as an assemble-to-order systems. Song (1998) estimates the order fill rate in such systems and discusses other examples from retailing.

2.3 Shelf Space Allocation Models

In some product segments such as grocery and pharmaceuticals, how much shelf space is allocated to a given product category is an important component of the assortment planning process. This view seems especially relevant for fast moving products whose demand is sufficiently high that a significant amount of inventory is carried on the shelf. This contrasts with other categories e.g., shoes, music, books where only one or two units are carried for most SKUs, hence amount of inventory and shelf space are not critical decisions at product level. As one example, Transworld Entertainment carries 50,000 SKUs in an average store but stock more than one of only the 300 best sellers.

In an influential paper Corstjens and Doyle (1981) suggest a method for allocating shelf space to categories. They perform store experiments to estimate sales of product i as $\alpha_i s_i^{\beta_i} \prod_j s_j^{\delta_{ij}}$, where s_i is the space allocated to product i, β_i is own space elasticity, and δ_{ij}s are the cross-space elasticities. Cost functions of the form $\gamma_i s_i^{\tau_i}$, are also estimated from the experiments. The problem of profit maximization with a shelf space constraint is solved within a geometric programming framework. Their results are significantly better than commercial algorithms that allocate space proportional to sales or to gross profit by ignoring interdependencies between product groups. The estimation and optimization procedures can not be applied to large problems, hence they elect to work with product groups rather than SKUs. Bultez and Naert (1988) apply the Corstjens and Doyle (1981) model at the brand level assuming symmetric cross elasticities (i.e., $\delta_{ij} = \delta$ for all i, j) within product groups. Their model is tested at four different Belgian supermarket chains, leading to encouraging results.

An interesting paper by Borin and Farris (1995) reports the sensitivity of the shelf space allocation models to forecast accuracy. They compare the solution with correct parameters to that with incorrect parameter estimates. Even when the error in parameter estimates are 24 %, the net loss in category return on inventory is just over 5 % compared to the optimal allocation based on true estimates. This proves the robustness of these models to estimation errors. Similar to these shelf space allocation papers, but using an inventory theoretic perspective, Urban (1998) models the own and cross product effects of displayed inventory on demand rate in a mathematical program and solves for shelf space allocation and optimal order-up-to quantities. He reports that on average a greedy heuristic yields solutions that are within 1 % of a solution obtained by genetic programming.

Irion et al. (2012) extend the Corstjens and Doyle model to study the shelf space allocation problem at the product level. Demand for each product is a function of its own and other products' shelf space through own and cross shelf space elasticities. The cost for each product consists of linear purchasing costs, inventory costs from an economic order quantity model, and a fixed cost of being included in the assortment. The objective is to allocate (integer) number of facings to each product in order to maximize profits under a total shelf space availability constraint and lower and upper bounds on the number of facings for each product. The problem is transformed into a mixed integer program (MIP) with linear constraints and objective function through a series of linearization steps. The linearization framework is general enough to accommodate several extensions. However, there is no empirical evidence that product level demand can be modeled as a function of the shelf space allocated to the product itself and competing products via own and cross space elasticities.

Shelf space allocation papers do not explicitly address assortment selection and inventory decisions and ignore the stochastic nature of demand.

2.4 Perception of Variety

Consumer choice models often assume that customers are perfectly knowledgeable about their preferences and the product offerings. Therefore, consumers are always better off when they choose from a broader set of products. However, empirical studies show that consumer choice is affected by their perception of the variety level rather than the real variety level. This perception can be influenced by the space devoted to a category, the presence or absence of a favorite item (Broniarczyk et al. 1998), or the arrangement of the assortment (Simonson 1999). Hoch et al. (1999) define a measure of the dissimilarity between product pairs as the count of attributes on which a product pair differs. They show that this measure is critical to the perception of variety of an assortment and that consumers are more satisfied with stores carrying those assortments perceived as offering high variety. van Herpen and Pieters (2002) find the impact of two attribute-based measures that significantly impact the perception of variety. These measures are entropy (whether all products have the same color or different colors) and dissociation between attributes (whether color and fabric choice across products are uncorrelated). The perception of variety at a store is especially important for variety-seeking consumers. Variety seeking consumers tend to switch away from the product consumed on the last occasion. Variety-seeking literature demonstrated that consumers adopt this behavior when purchasing food or choosing among hedonic products such as restaurants and music. See Kahn (1995) for a review. Intrapersonal factors (e.g., satiation and the need for stimulation), external factors (e.g., price change, new product introduction), and uncertainty about future preferences promote variety-seeking behavior. On a final note, variety can even negatively affect consumers experience: confusion or complexity due to higher variety may cause dissatisfaction of consumers and decrease sales (Huffman and Kahn 1998).

3 Demand Models

This section provides a review of demand models as background for assortment planning models. We first present the empirical evidence for consumer driven substitution which is a fundamental assumption in many assortment planning models. The Multinomial Logit model is a discrete consumer choice model, which assumes that consumers are rational utility maximizers and derive customer choice behavior from first principles. Exogenous demand models directly specify the demand for each product and what an individual does when the product he or she demands is not available. The locational choice model is also a utility-based model. Before proceeding, we will define the notation for assortment planning in a single subcategory at a single store. This notation is common throughout this chapter and additional time or store subscripts are introduced when necessary.

N The set of products in a subcategory, $N = \{1, 2, .., n\}$,
S The subset of products carried by the retailer, $S \subset N$,
r_j Selling price of product j,
c_j Purchasing cost of product j,
λ Mean number of customers visiting the store per period.

3.1 Consumer Driven Substitution

We define two types of substitution with a supply side view of the causes of substitution: *Stockout-based* substitution is the switch to an available variant by a consumer when her favorite product is carried in the store, but is stocked-out at the time of her shopping. *Assortment-based* substitution is the switch to an available variant by a consumer when her favorite product is not carried in the store.

The substitution possibilities in retailing can be classified into three groups. (a) Consumer shops a store repeatedly for a daily consumable, and one day she finds it stocked out so she buys another. This is an example of stockout-based substitution. (b) Consumer has a favorite product based on ads or her past purchases at other stores, but the particular store she visited on a given day may not carry that product. This is an example of assortment-based substitution. (c) Consumer chooses her favorite from what she sees on the shelf and buys it if it is better than her no purchase option. In this case, there may be other products she may have preferred, but she didn't see them either because the retailer didn't carry them or they are stocked out. This could be an example for either substitution type depending on whether the first choice product is temporarily stocked out or not carried at that store. First two cases fit repeat purchases like food and the third fits one time purchases like apparel.

Let's focus on the options of a consumer who can not find her favorite product in a store, because it is either temporarily stocked out or not carried at all. She can (a) buy one of the available items from that category (substitute), (b) decide to come back later for that product (delay), (c) decide to shop at another store (lost customer). If the consumer chooses to substitute, the sale is lost from the perspective of the first favorite product. Table 8.1 summarizes the findings of empirical studies on the consumer response to stockouts. The most recent one, Gruen et al. (2002) examine consumer response to stockouts across eight categories at

Table 8.1 Consumer response to stockouts in six studies of substitute-delay-leave behavior

	Substitute (%)	Delay (%)	Leave (%)
Progressive Grocer (1968a,b)	48	24	28
Walter and Grabner (1975)	83	3	14
Schary and Christopher (1979)	22	30	48
Emmelhainz et al. (1991)	36	25	39
Zinn and Liu (2001)	62	15	23
Gruen et al. (2002)	45	15	40

retailers worldwide and report that 45 % of customers substitute, i.e., buy one of the available items from that category, 15 % delay purchase, 31 % switch to another store, and 9 % never buy that item.

The above mentioned papers study the consumer response to stockouts, i.e. stockout based substitution, although none of them explicitly excludes assortment-based substitution. Campo et al. (2004) investigate the consumer response to out-of-stocks (OOS) as opposed to permanent assortment reductions (PAR). They report that although the retailer losses in case of a PAR may be larger than those in case of an OOS, there are also significant similarities in consumer reactions in the two cases and OOS reactions for an item can be indicative of PAR responses for that item.

3.2 Multinomial Logit

The Multinomial Logit (MNL) model is a utility-based model that is commonly used in economics and marketing literatures. We create product 0 to represent the no-purchase option, i.e., a customer that chooses 0 does not purchase any products. Each customer visiting the store associates a utility U_j with each option $j \in S \cup \{0\}$. The utility is decomposed into two parts, the deterministic component of the utility u_j and a random component ε_j.

$$U_j = u_j + \varepsilon_j.$$

The random component is modeled as a Gumbel random variable. Also known as Double Exponential or Extreme value Type-I, it is characterized by the distribution

$$Pr\{X \leq \varepsilon\} = \exp(-\exp - (\varepsilon/\mu + \gamma))),$$

where γ is Euler's constant (0.57722). Its mean is zero, and variance is $\mu^2\pi^2/6$. A higher μ implies a higher degree of heterogeneity among the customers. The realizations of ε_j are independent across consumers. Therefore, while each consumer has the same expected utility for each product, realized utility may be different. This can be due to the heterogeneity of preferences across customers or unobservable factors in the utility of the product to the individual.

An individual chooses the product with the highest utility among the set of available choices. Hence, the probability that an individual chooses product j from $S \cup \{0\}$ is

$$p_j(S) = \Pr\left\{ U_j = \max_{k \in S \cup \{0\}} (U_k) \right\}.$$

The Gumbel distribution is closed under maximization. Using this property, we can show that the probability that a customer chooses product j from $S \cup \{0\}$ is

$$p_j(S) = \frac{e^{u_j/\mu}}{\sum\limits_{k \in S \cup \{0\}} e^{u_k/\mu}}. \tag{8.1}$$

See Anderson et al. (1992) for a proof. This closed form expression makes the MNL model an ideal candidate to model consumer choice in analytical studies. See Ben-Akiva and Lerman (1985) for applications to the travel industry, Anderson et al. (1992) for MNL based models of product differentiation, Basuroy and Nguyen (1998) for equilibrium analysis of market share games and industry structure. Moreover, starting with Guadagni and Little (1983), marketing researchers found that MNL model is very useful in estimating demand for a group of products. We will briefly discuss the parameter estimation of MNL model in Sect. 5.1. For more details on the MNL model and its relation to other choice models, see Anderson et al. (1992) or Mahajan and van Ryzin (1999).

The major criticism of the MNL model stems from its Independence of Irrelevant Alternatives (IIA) property. This property holds if the ratio of choice probabilities of two alternatives is independent of the other alternatives in the choice process. Formally, this property is

for all $R \subset N, T \subset N, R \subset T$, *for all* $j \in R, k \in R$,

$$\frac{p_j(R)}{p_k(R)} = \frac{p_j(T)}{p_k(T)}.$$

IIA property would not hold in cases where there are subgroups of products in the choice set such that the products within the subgroup are more similar with each other than across subgroups. Consider an assortment with two products from different brands. If brand loyalty is high, adding a new product from the first brand can cannibalize the sales of its sister product more than the rival product. IIA does not capture this important aspect of consumer choice. Another example that illustrates this property is the "blue bus/red bus paradox": Consider an individual going to work and has the same probability of using his or her car or of taking the bus: $\Pr\{car\} = \Pr\{bus\} = 1/2$. Suppose now that there are two buses available that are identical except for their color, red or blue. Assume that the individual is indifferent about the color of the bus he or she takes. The choice set is {car, red bus, blue bus}. One would intuitively expect that $\Pr\{car\} = 1/2$ and $\Pr\{red\ bus\} = \Pr\{blue\ bus\} = 1/4$. However, the MNL model implies that $\Pr\{car\} = \Pr\{red\ bus\} = \Pr\{blue\ bus\} = 1/3$.

The Nested Logit Model introduced by Ben-Akiva and Lerman (1985) is one way to deal with the IIA property. A two-stage nested process is used for modeling choice, e.g., first brand choice then SKU choice. The choice set N is partitioned into subsets N_l, $l = 1, .., m$ such that $\cup_{l=1}^{m} N_l = N$ and $N_l \cap N_k = \emptyset$ for any l and k. The individual chooses with a certain probability one of the subsets, from which he or she chooses a variant from that subset. The utility from the choice within subset N_l is also Gumbel distributed with mean $\mu \ln \sum_{j \in N_l} e^{u_j/\mu}$ and scale parameter μ. As a result, the choice process between the subsets follows the MNL model as well and the probability that a consumer chooses variant j in subset N_l is

$$P_j(N) = P_{N_l}(N) * P_j(N_l).$$

Chapter 2 in Anderson et al. describes the Nested Logit in great detail. In the Nested Logit Model, the IIA property no longer holds when two alternatives are not in the same subgroup. However, the use of the Nested Logit requires the knowledge of key attributes and their hierarchy for consumers and makes estimation problems more difficult. Nested Logit model is used in modeling the competition between two-multiproduct firms in several studies (Anderson et al. 1992; Cachon et al. 2008).

Another related shortcoming of the MNL model is related to substitution between different products. The MNL model in its simplest form is unable to capture an important characteristic of the substitution behavior. The utility of the no-purchase option with respect to the utility of the products in S determines the rate of substitution. Consider the following example, where $S = \{1, 2\}, \mu = 1,$ and $u_0 = u_1 = u_2$. The share of each option is determined by the implication of MNL that the probability of choosing option i is $\exp(u_i)/(\exp(u_0) + \exp(u_1) + \exp(u_2)) = 1/3$ for $i = 0, 1, 2$. Hence, two thirds of the customers are willing to make a purchase from the category. If the second product is unavailable, the probability of her choosing the first product is $\exp(u_1)/(\exp(u_0) + \exp(u_1)) = 1/2$. That is, half of the consumers whose favorite is stocked out will switch to the other product as a substitute and the other half will prefer no-purchase alternative to the other product. In this example, the penetration to the category (purchase incidence) is 2/3 and the average substitution rate is 1/2. These two quantities are linked via u_i's. We can control the substitution rate by varying u_0, but that also determines the initial penetration rate to the category. Hence, it is not possible with this model to have two categories with the same penetration rate but different substitution rates, which we have found severely limits the applicability of this model.

Miranda Bront et al. (2009) show that the CDLP model of the assortment problem with multiple segments is NP-hard and propose a column generation algorithm. Rusmevichientong and Topaloglu (2012) propose a robust formulation of the assortment optimization problem.

3.3 Exogenous Demand Model

Exogenous demand models directly specify the demand for each product and what an individual does when the product he or she demands is not available. There is no underlying consumer behavior such as a utility model that generates the demand levels or that explains why consumers behave as described in the model. As mentioned before, this is the most commonly used demand model in the literature on inventory management for substitutable products. The following assumptions fully characterize the choice behavior of customers.

(A1) Every customer chooses her favorite variant from the set N. The probability that a customer chooses product j is denoted by p_j. $\sum_{j \in N \cup \{0\}} p_j = 1$.

(A2) If the favorite product is not available for any reason, with probability δ she chooses a second favorite and with probability $1 - \delta$ she elects not to purchase. The probability of substituting product j for k is α_{kj}.

When the substitute item is unavailable, consumers repeat the same procedure: decide whether or not to purchase and choose a substitute. The lost sales probability $(1 - \delta)$ and the substitution probabilities could remain the same for each repeated attempt or specified differently for each round.

As a result of (A1) average demand rate for product j is $d_j = \lambda\, p_j$, and total demand to the category is $\sum_{j \in N} d_j = \lambda(1 - p_0)$.

α_{kj} is specified by a substitution probability matrix that can take different forms to represent different probabilistic mechanisms. Consider the following examples for a four-product category.

Random substitution matrix

$$\begin{bmatrix} 0 & \frac{\delta}{n-1} & \frac{\delta}{n-1} & \frac{\delta}{n-1} \\ \frac{\delta}{n-1} & 0 & \frac{\delta}{n-1} & \frac{\delta}{n-1} \\ \frac{\delta}{n-1} & \frac{\delta}{n-1} & 0 & \frac{\delta}{n-1} \\ \frac{\delta}{n-1} & \frac{\delta}{n-1} & \frac{\delta}{n-1} & 0 \end{bmatrix}$$

Adjacent substitution matrix

$$\begin{bmatrix} 0 & \delta & 0 & 0 \\ \delta/2 & 0 & \delta/2 & 0 \\ 0 & \delta/2 & 0 & \delta/2 \\ 0 & 0 & \delta & 0 \end{bmatrix}$$

Within subgroups substitution matrix

$$\begin{bmatrix} 0 & \delta & 0 & 0 \\ \delta & 0 & 0 & 0 \\ 0 & 0 & 0 & \delta \\ 0 & 0 & \delta & 0 \end{bmatrix}$$

Proportional substitution matrix

$$\begin{bmatrix} 0 & \delta d_2/(\lambda - d_1) & \delta d_3/(\lambda - d_1) & \delta d_4/(\lambda - d_1) \\ \delta d_1/(\lambda - d_2) & 0 & \delta d_3/(\lambda - d_2) & \delta d_4/(\lambda - d_2) \\ \delta d_1/(\lambda - d_3) & \delta d_2/(\lambda - d_3) & 0 & \delta d_4/(\lambda - d_3) \\ \delta d_1/(\lambda - d_4) & \delta d_2/(\lambda - d_4) & \delta d_3/(\lambda - d_4) & 0 \end{bmatrix}$$

The single parameter δ enables us to differentiate between product categories with low and high substitution rates. The adjacent substitution matrix assumes that products are ordered along an attribute space and allows for substitution between neighboring products only. For example, if a customer can't find 1 % milk in stock, she may be willing to accept either 2 % or skim, but not whole milk. Subgroups substitution matrix allows for substitution within the subgroups only. For example, in the coffee category, consumers may treat decaffeinated coffee and regular coffee as subgroups and not substitute between subgroups.

In the proportional substitution model, the general expression for α_{kj} is

$$\alpha_{kj} = \delta \frac{d_j}{\sum_{l \in N \setminus \{k\}} d_l}. \tag{8.2}$$

The proportional substitution matrix has properties that are consistent with what would happen in a utility-based framework such as the MNL model. $\alpha_{kj} > \alpha_{kl}$ if $d_j > d_l$. Suppose that a store doesn't carry the whole assortment, i.e., $N \setminus S \neq \emptyset$. Since only one round of substitution is allowed, the realized substitution rate from variant k to other products is $\sum\limits_{j \in S} \alpha_{kj} = \delta \sum\limits_{j \in S} d_j / \sum\limits_{l \in N \setminus \{k\}} d_l$, which is increasing in the set S. This means that a consumer who can not find her favorite variant in the store is more likely to buy a substitute, as the set of potential substitutes grows.

We next state an assumption commonly made in assortment planning models for tractability.

(A3) No more attempts to substitute occur. Either the substitute product is available and the sale is made, or the sale is lost.

Limiting the number of substitution attempts (A3) is not too restrictive. Smith and Agrawal (2000) show that number of attempts allowed has a smaller effect as more items are stocked, because the probability of finding a satisfactory item by the second try quickly approaches one. Kök (2003) presents an example where effective demands under a three-attempts substitution model with rate $\delta = 0.5$ can be approximated almost perfectly with a single-attempt-substitution model with rate $\delta = 0.58$.

The exogenous demand model has more degrees of freedom than the MNL model. Since the options in the choice set are assumed to be homogenous, MNL model is unable to capture the types of adjacent substitution, one-product substitution, or within subgroup substitution. In the MNL model the substitution rates depend on the relative utility of the options in $N \cup \{0\}$. This is both an advantage and a disadvantage for the MNL model. The advantage is that it allows one to easily incorporate marketing variables such as prices and promotions into the choice model. The disadvantage is that it cannot differentiate between the initial choice and substitution behavior. Unlike the MNL model, the exogenous demand model can differentiate between categories that have same initial demand for the category but different substitution rates through the choice of p_0 and δ. Therefore, the MNL model cannot treat assortment-based and stockout-based substitutions differently. In contrast, it is certainly possible to use a different δ or different substitution probability matrices for assortment-based and stockout-based substitutions in the exogenous demand model.

3.4 Locational Choice Model

Also known as the address or the characteristics approach, the locational choice model was originally developed by Hotelling (1929) to study the pricing and location decisions of competing firms. Extending Hotelling's work, Lancaster (1966, 1975) proposed a locational model of consumer choice behavior. In this model, products are viewed as a bundle of their characteristics (attributes) and each product can be represented as a vector in the characteristics space, whose components indicate how much of each characteristic is embodied in that product. For example, defining characteristics of a car include its engine size, gas consumption, and reliability. Each individual is characterized by an ideal point in the characteristics space, which corresponds to his or her most preferred combination of characteristics.

Suppose that there are m characteristics of a product. Let z_j denote the location of variant j in R^m. Consider a consumer whose ideal product is defined by $y \in R^m$. The utility of variant j to the consumer is

$$U_j = k - r_j - g(y, z_j),$$

where k is a positive constant, r_j is the price, and $g : R^m \to R$ is a distance function, representing the disutility associated with the distance from the consumer's ideal point, e.g., Euclidean distance or the rectilinear distance. The consumer chooses the variant that gives him or her the maximum utility. For an extensive discussion of the address approach and its relation to stochastic utility models such as the MNL model, the reader is referred to Chap. 4 in Anderson et al. (1992).

There is one major difference between the locational choice model and the MNL model. In the MNL model, substitution can happen between any two products. In the locational choice model however, IIA property does not hold and substitution between products is localized to products with specifications that are close to each other in the characteristics space. Hence, the firm can control the rate of substitution between products by selecting their locations to be far apart or close to each other.

4 Assortment Selection and Inventory Planning

The majority of the papers focus on assortment decisions at a single store. Most papers take a static view of the assortment planning problem, that is the assortment decisions are made once and inventory costs are computed either from a single period model or the steady-state average of a multi-period model. In Sects. 4.1–4.3, we review four such papers categorized according to the demand model that they are based on. The papers based on the choice models are more stylized but are able to obtain structural properties of the optimal solution. The papers based on the exogenous demand model are more flexible and have more applicability because they allow for more realistic details in modeling, such as nonidentical prices and case packs. In Sect. 4.4, we review assortment planning papers with supply chain considerations. Section 4.5 discusses a dynamic assortment planning model in which the retailer has a chance to update its assortment throughout the season as it updates its demand estimates every period for products in the assortment. A recent development in the assortment planning literature is the consideration of multiple categories, where consumers are basket shoppers and the assortment decisions across categories are interdependent. In Sect. 4.7, we discuss two such papers. The first presents an optimization method and the second discusses the long-run impact of variety by considering store choice decisions of consumers.

4.1 Assortment Planning with Multinomial Logit: The van Ryzin and Mahajan Model

van Ryzin and Mahajan (1999) formulate the assortment planning problem by using a MNL model of consumer choice. Assume $r_j = r$ and $c_j = c$ for all j. Products are indexed in descending order of their popularity, i.e., such that $u_1 \geq u_2 \geq .. \geq u_n$. Define $v_j = e^{u_j/\mu}$. By the MNL share formula, the probability that a customer demands product j is

$$p_j(S) = \frac{v_j}{\sum\limits_{k \in S \cup \{0\}} v_j}. \tag{8.3}$$

We assume consumers make their product choice (if any) when they observe the assortment, and they do not look for a substitute if the product of their choice is stocked out. Hence, $p_j(S)$ is independent of the inventory status of the products in S. Note that the demand increase in product j due to the decision $S \subseteq N$ is

$$p_j(S) - p_j(N).$$

This demand increase is due to what is termed assortment-based substitution and is comprised of demand from consumer who would have preferred a product in $N - S$ but had to substitute to product j. van Ryzin and Mahajan (1999) also calls this static substitution.

In contrast, in dynamic substitution, consumers observe the inventory levels of all products at the time of their arrival and make their product choice among the products that are available. Hence, dynamic substitution includes both assortment- and stockout-based substitution.

The expected profit of a variant $j \in S$ is

$$\pi_j(S) = (r - c)\lambda p_j(S) - C(\lambda p_j(S)),$$

where $C(\cdot)$ is the operational costs. The cost function is assumed to be concave and increasing to reflect the economies of scale in inventory models such as the EOQ or the newsvendor models.

The objective is to maximize the total category profits by solving

$$\max_{S \subset N} \sum_{j \in S} \pi_j(S).$$

The optimal assortment finds a balance between including a new product and increasing the total demand to the category and cannibalizing the demand of other products' sales and increasing their average cost.

Consider the net profit impact of adding a variant j to assortment S. Define $S_j = S \cup \{j\}$.

$$h(v_j) = \pi_j(S_j) - \left(\sum_{k \in S} \pi_k(S) - \sum_{k \in S} \pi_k(S_j) \right)$$

If the profit of product j is more than the sum of the profit losses of the products in S, then adding j improves profits.

Theorem 1 *The function h(v_j) is quasi-convex in v_j in the interval [0,∞).*

Since a quasi-convex function achieves its maximum at the end points of the interval, the profit is maximized either by not adding a product to the assortment or by adding the product with the highest v (i.e., the most popular product). This observation leads to the following result that characterizes the structure of the optimal assortment. Define the popular assortment set:

$$P = \{\{\}, \{1\}, \{1,2\}, .., \{1,2,..,n\}\}.$$

Theorem 2 *The optimal assortment is always in the popular assortment set.*

This result is intuitive and powerful: it reduces the number of assortments to be considered from 2^n to n. Since only assortment-based substitution is considered, the demand for each product, the optimal inventory level and the resulting profit can be computed for each of the n assortments in the popular assortment set. The above theorems as stated are from Cachon et al. (2005). van Ryzin and Mahajan (1999) originally proved this result for a cost function from the newsvendor model. Specifically, they use the expected costs of a newsvendor model assuming that D is distributed according to a Normal distribution with mean λ and standard deviation σ. The optimal stocking level of product j is the newsvendor stocking quantity:

$$x_j = \lambda p_j(S) + z\sigma\left(\lambda p_j(S)\right)^{\beta},$$

where $z = \Phi^{-1}(1 - c/r)$ and $\beta \in [0,1)$ controls the coefficient of variation of the demand to product j as a function of its mean. The resulting cost function is

$$C(\lambda p_j(S)) = r\sigma \frac{e^{-z^2}}{\sqrt{2\pi}} \left(\lambda p_j(S)\right)^{\beta}.$$

The authors show that a deeper assortment is more profitable with a sufficiently high price, and a sufficiently high no-purchase preference. In order to compare different merchandising categories, the authors define the fashion of a category using majorization arguments. In a more fashionable category, the utility across products are more balanced, therefore in expectation the market shares of all products are evenly distributed. The paper shows that everything else being equal, the profit of a more fashionable category is lower due to the fragmentation of demand.

This model captures the main trade-off between variety and the increased average inventory costs. The analysis leads to the elegant results that establish the structural properties of the optimal assortment. However, not all assortment planning problems fit the assumption of homogenous group of products with identical prices and costs. The style/color/size combination of shirts in a clothing retailer may be a good example. Even then, the substitutions would occur across styles/

colors but not sizes. The assumption that there is a single opportunity to make assortment and inventory decisions can be defended in products with short life cycles, where the season is too short to make changes in the assortment and bring the new products to market before the season is over. Clearly, the main result (Theorem 2) does not hold when products have nonidentical price, cost parameters, or different operational characteristics such as demand variance, case pack, and minimum order quantity.

4.1.1 Extensions

Mahajan and van Ryzin (2001a) study the same problem under dynamic substitution. That is, the retailer faces the problem of finding the optimal product selection and stocking levels where customers dynamically substitute among products when inventory is depleted. Consider a customer with the following realization of the utilities: $u_6 > u_4 > u_3 > u_5 > u_0 > u_1 > u_2$. Suppose that the store carries assortment $S = \{1,2,3,4\}$. In the static substitution model, this consumer would choose product 4, buy it if it is available and leave the store if it is not. In the dynamic substitution model, products 4, 3, and 5 are all acceptable to the customer, in that order of preference. Depending on the inventory levels of those products, she will buy the one that is available in the store at the time she visited the store, and won't buy anything only if none of those three products is available. Using a sample path analysis, the authors show that the problem is not even quasi-concave. By comparing the results of a stochastic gradient algorithm with two newsvendor heuristics, they conclude that the retailer should stock more of the more popular variants and less of the less popular variants than a traditional newsvendor analysis suggests. Also, the numerical results support the theoretical insight (Theorem 2) obtained under static substitution. Maddah and Bish (2004) extend the van Ryzin Mahajan model by considering the pricing decisions as well.

Cachon et al. (2005) study the van Ryzin and Mahajan (1999) model in the presence of consumer search, motivated by the following observation: Even when a consumer finds an acceptable product at the retail store, the consumer still faces an uncertainty about the products outside the store's assortment. Therefore, she may be willing to go to another store and explore other alternatives with the hope of finding a better product. In the independent search model, consumers expect each retailer's assortment to be unique, and hence utility of search is independent of the assortment. Examples for this setting include jewelry stores and antique dealers. In the overlapping assortment search model, products across retailers overlap, hence the value of search decreases with the assortment size at the retailer. For example, all retailers choose their digital camera assortments from the product lines of a few manufacturers. In contrast to the no-search model, in the presence of consumer search it may be optimal to include an unprofitable product in the assortment. Therefore, failing to incorporate consumer search in assortment planning results in narrower assortments and lower profits.

Vaidyanathan and Fisher (2012) study the assortment planning problem under a setup similar to van Ryzin and Mahajan (1999), but in the presence of more general demand distributions. They approximate the expected profit function and evaluate simple heuristics to select the optimal assortment and set inventory levels, in the presence of stock-out substitution. They also present analytical bounds on the error due to optimizing an approximate profit function instead of the exact one.

Miller et al. (2010) consider the retailer's assortment selection problem with heterogeneous customers and test the impact of different consumer choice models on the optimal assortment. They develop a sequential choice model in which customers first form Consideration Sets and then make product choices based on the MNL model.

Li (2007) extend the van Ryzin and Mahajan (1999) and show that under continuous traffic, the optimal assortment consists of a set of products with the highest profit rates, even when product margins are unequal.

Kök and Xu (2011) use a nested logit model to study assortment decisions for a product category with heterogeneous product types from two brands. They consider two different hierarchical structures for the nests: a brand-primary model in which consumers choose a brand first, then a product type in the chosen brand, and a type-primary model in which consumers choose a product type first, and then a brand within that product type. They extend the structural properties of assortment decisions characterized by van Ryzin and Mahajan (1999) to the case of Nested Logit. A more detailed discussion of this paper can be found in Sect. 4.6.

Alptekinoğlu and Grasas (2014) apply the nested logit model to study assortment decisions under consumer returns, for a set of horizontally differentiated products. They show that when refund amounts are sufficiently high, or when returns are disallowed, the optimal assortment consists of only the most popular products, a result consistent with van Ryzin and Mahajan (1999). However, when return policies are relatively strict, and refund amounts are low, they find that it might be optimal for the retailer to offer a mix of the most popular and eccentric products. They support their findings with some empirical evidence that eccentric products are usually associated with higher return probabilities.

Davis et al. (2014) show that the assortment optimization problem under the nested logit model can be solved in polynomial time, when customers are assumed to always make their purchase from the selected nest, and the nest dissimilarity parameters satisfy certain conditions. In the absence of either of these assumptions, they demonstrate that the problem is NP-hard.

Alptekinoğlu and Semple (2013) propose a new discrete choice model, termed the Exponomial Choice Model, which modifies the MNL model, by assuming exponentially distributed random errors. They obtain closed form expressions for the choice probabilities and find that unlike the MNL model, the exponomial model does not suffer from the independence of irrelevant alternatives (IIA) property. Additionally, they show that the exponomial choice model is easy to estimate, since the loglikelihood function is concave in the unknown parameters. They derive structural properties of the optimal assortment and prices, under a number of scenarios. Finally, they estimate the exponomial choice model on two sets of choice data and compare the results with the MNL model.

4.1.2 Preference Ordering Models

Honhon et al. (2010) study a single-period joint assortment and inventory planning problem when customers are classified based on their preference ordering of products. They assume that total customer demand is random, and the market is comprised of fixed proportions of different customer types, based on preference ordering. They develop efficient, pseudopolynomial time algorithms to solve the resulting assortment optimization problem.

Honhon et al. (2012) study the optimal assortment problem under the assumption that (a) customers can be characterized into types based on a rank-ordered list of products they are willing to purchase, (b) proportion of consumers of each type is random and (c) purchases are dynamic, consumer-driven and stockout based. Following Honhon et al. (2010), the authors relax the assumption of random proportions to show that the expected profits for the resulting fixed proportions model (FP) can be used to construct tight bounds on the expected profits for the random proportions model. Finally, they use these bounds and numerical simulations to (a) study optimality gap as a function of problem parameters and (b) conclude that the FP heuristic performs favorably to other previously known heuristics in literature.

Honhon et al. (2012) study the optimal assortment selection problem under four different ranking-based consumer choice models, the one-way substitution, the locational choice, the outtree, and the intree preference model. They model the problem assuming that the retailer incurs a fixed carrying cost for every product offered, a goodwill penalty when a customer is unable to find his first choice, and lost sales penalty when a customer is not able to find any acceptable product. Under these assumptions, they find that the first three models can be solved efficiently using a shortest path algorithm or dynamic program. For the intree preference model, they construct an algorithm that is efficient and performs better than enumeration based methods in numerical experiments.

Pan and Honhon (2012) study the assortment planning problem for a category of vertically differentiated products. There is a fixed cost to include a product in the assortment and additional variable costs are incurred per unit sold. Customers are utility maximizers and differ in their valuation of quality, which is exogenously determined. They find that under fixed selling prices, the optimal assortment might include strictly dominated products, that are less attractive on every possible dimension, as compared to at least one other product not carried in the assortment. In the scenario where the retailer can set the selling prices, they find that this counter-intuitive feature of the optimal assortment disappears. They propose several efficient algorithms to determine the optimal assortment and pricing structure, and test them on real data for two product categories.

4.2 Assortment Planning Under Exogenous Demand Models

In this subsection, we review two closely related assortment planning models that consider both assortment-based and stockout-based substitution. Smith and Agrawal (2000) focus on constructing lower and upper bounds to the problem in order to formulate a mathematical program. Kök and Fisher (2007) formulate the problem in the context of an application at a supermarket chain and proposes a heuristic solution to a similar mathematical program. They also provide structural results on the assortments that generate new insights and guidelines for practitioners and researchers.

4.2.1 Smith and Agrawal Model

Smith and Agrawal (2000) (hereafter SA) study the assortment planning problem with the exogenous demand model. SA models the arrival process of customers carefully and updates the inventory levels after each customer arrival. Given assortment S, SA sets the stocking level of each product to achieve exogenously determined service levels f_j. Let $g_j(S,m)$ denote the probability that m^{th} customer chooses product j and $A_k(S,m)$ a binary variable indicating the availability of product k when the m^{th} customer arrived. Both clearly depend on the choice of previous customers and the number of substitution attempts made by the customer. For one substitution-attempt-only model,

$$g_j(S,m) = d_j + \sum_{k \notin S} d_k \alpha_{kj} + \sum_{k \in S \setminus \{j\}} d_k \alpha_{kj} (1 - A_k(S,m))$$

The first term is the original demand for product j, the second term is the demand from assortment substitution and the third from stockout substitution. Since exactly determining $g_j(S,m)$ is complex, SA develops lower and upper bounds. The lower bound is achieved by considering only assortment-based substitution and the upper bound by assuming that products achieve f_j in-stock probability even for the first customer, hence overestimating stockout substitution. Specifically,

$$\begin{aligned} & h_j(S) \le g_j(S,m) \le H_j(S) \quad \text{for all} \quad m, \text{where} \\ & h_j(S) = d_j + \sum_{k \notin S} d_k \alpha_{kj}, \\ & H_j(S) = d_j + \sum_{k \notin S} d_k \alpha_{kj} + \sum_{k \in S \setminus \{j\}} d_k \alpha_{kj} f_k. \end{aligned} \tag{8.4}$$

SA shows that these bounds are tight and uses the lower bound $h_j(S)$ to approximate the demand rate. That is, effective demand for product j given assortment S follows a distribution with mean $h_j(S)$. SA provides similar bounds to the demand rate under the repeated-attempts substitution model. Agrawal and Smith (1996) found that

Negative Binomial distribution (NBD) fits retail sales data very well. SA shows that when the total number of customers that visit a store is distributed with NBD, the demand for each product would also follow NBD.

The optimization problem is to maximize the total category profits:

$$\max_{S\subset N} Z = \sum_{j\in S} \pi_j(S)$$

where the profit function for each product j is the newsvendor profit minus the fixed cost of stocking an item V_j.

$$\pi_j(S) = \left(r_j - c_j\right)h_j(S) - c_j E[x_j - D_j|h_j(S)]^+ - \left(r_j - c_j\right)E[D_j - x_j|h_j(S)]^+ - V_j,$$

where D_j is the random variable representing the demand for product j, x_j is the optimal newsvendor stocking quantity to achieve the target stocking level $f_j = 1 - c_j/r_j$, e.g., $\Pr\{D_j \geq x_j \,|\, h_j(S)\} = f_j$ for a continuous demand distribution. Incorporating salvage value, or holding costs to the newsvendor profit function above is trivial.

This optimization problem is a nonlinear integer programming problem. SA proposes solving the problem via enumeration for small n and a linearization approximation for large n. A single constraint such as a shelf space or a budget constraint can be incorporated into the optimization model. SA proposes a Lagrangian Relaxation approach followed by a one-dimensional search on the dual variable for the resulting mathematical program.

Several insights are obtained from illustrative examples. Substitution effects reduce the optimal assortment size when fixed costs are present. However, even when there are no fixed costs present, substitution effects can reduce the optimal assortment size, because products have different margins. Contrary to the main result of van Ryzin and Mahajan (1999), it may not be optimal to stock the most popular item—a result of the adjacent substitution matrix or the one-item substitution matrix.

4.2.2 Kök and Fisher Model

The methodology described in Kök and Fisher (2007) is applied at Albert Heijn, BV, a leading supermarket chain in the Netherlands with 1,187 stores and about $10 billion in sales. The replenishment system at Albert Heijn is typical in the grocery industry. All the products in a category are subject to the same delivery schedule and fixed leadtime. There is no backroom, therefore orders are directly delivered to the shelves. Shelves are divided into *facings*. SKUs in a category share the same shelf area but not the same facing, i.e., only one kind of SKU can be put in a facing. Capacity of a facing depends on the depth of the shelf and the physical size of a unit of the SKU. The inventory model is a periodic review model with stochastic demand, lost sales and positive constant delivery lead-time. The number of facings allocated to product j, f_j, determines its maximum level of inventory, $k_j f_j$, where k_j is the capacity of a facing. At the beginning of each period, an integral number of case packs (batches) of size b_j is ordered to take the inventory position as close as possible to the maximum inventory level without exceeding it. Case sizes vary significantly across products and significantly affect returns from inventory. The performance measure is gross profit, which is per-unit margin times sales minus selling price times disposed inventory.

We focus on a single subcategory of products initially for expositional simplicity and then explain how to incorporate the interactions between multiple subcategories. The decision process involves allocating a discrete number of facings to each product in order to maximize total expected gross profits subject to a shelf space constraint:

$$\begin{aligned} &\max_{f_j, j\in N} Z(\mathbf{f}) = \sum_j G_j(f_j, D_j(\mathbf{f},\mathbf{d})) \\ &s.t. \qquad \sum_j f_j w_j \le ShelfSpace_{AP} \\ &\qquad\qquad f_j \in \{0,1,2,..\}, \quad \text{for all } \; j \end{aligned} \tag{8.5}$$

where f_j is the number of facings allocated to product j, and w_j is the width of a facing of product j. G_j is the (long run) average gross profit from product j given f_j and demand rate D_j. Due to substitution, effective demand for a product includes the original demand for the product and substitution demand from other products. Hence, $D_j(\mathbf{f},\mathbf{d})$, the effective demand rate of product j, depends on the facing allocation and the demand rates of all products in the subcategory, i.e., $\mathbf{f} = (f_1, f_2, .., f_n)$ and $\mathbf{d} = (d_1, d_2, .., d_n)$, where d_j is the original demand rate of product j (i.e., number of customers who would select j as their first choice if presented with all products in N). The store's assortment is denoted S and is determined by the facing allocation, i.e., $S = \{j \in N : f_j > 0\}$.

Similar to SA, the effective demand rate function under this substitution model is

$$D_j(\mathbf{f},\mathbf{d}) = d_j + \left(\sum_{k:f_k=0} \alpha_{kj} d_k + \sum_{k:f_k>0} \alpha_{kj} L_k(f_k, d_k) \right) \tag{8.6}$$

where the L_k function is the lost sales (average unmet demand) of product k. In our application we estimate $L_k(f_k, d_k)$ via simulation. In (8.6), $\sum_{k:f_k=0} \alpha_{kj} d_k$ is the demand for j due to assortment-based substitution and $\sum_{k:f_k>0} \alpha_{kj} L_k(d_k, f_k)$ is the demand for j due to stockout-based substitution.

In a stochastic inventory model as described above, G_j is a nonlinear function of the allocated facings to product j. It is a function of the facings of product j (f_j), and the facings of all other SKUs in a subcategory through the D_j function. Hence, AP is a knapsack problem with a nonlinear and nonseparable objective function, whose coefficients need to be calculated for every combination of the decision variables. Even if we rule out stockout-based substitution, we need to consider 'in' and 'out' of the assortment values for all products leading to 2^n combinations.

We propose the following iterative heuristic that solves a series of separable problems. The details of the algorithm can be found in Kök and Fisher (2007). We set $D_j(\mathbf{f}, \mathbf{d}) = d_j$ for all j and solve (AP) with the original demand rates resulting in a particular facings allocation $\mathbf{f^0}$. At iteration t, we recompute $D_j(\mathbf{f}^{t-1}, \mathbf{d})$ given δ for all j according to Eq. (8.6). Note that $\sum_j G_j\left(f_j^t, D_j(\mathbf{f}^{t-1}, \mathbf{d})\right)$ is separable now, because $D_j(\mathbf{f}^{t-1}, \mathbf{d})$ are computed a priori. We then solve (AP) with $Z(\mathbf{f}^t) = \sum_j G_j\left(f_j^t, D_j(\mathbf{f}^{t-1}, \mathbf{d})\right)$ via a Greedy Heuristic. We keep iterating until f_j^t converges for all j. In a computational study, the Iterative Heuristic performs very well with an average optimality gap of 0.5 %.

(AP) can be generalized to multiple subcategories of products that share the same shelf space by including several subcategories in the summations in the objective function and the shelf space constraint. Let subscript $i = 1, .., I$ be the subcategory index. The objective function in the multiple subcategory case would be $Z(\mathbf{f}) = \sum_i \sum_j G_{ij}(f_{ij}, D_{ij}(\mathbf{f}_i, \mathbf{d}_i))$, the shelf space constraint can be modified similarly.

Structural Properties of the Iterative Heuristic

The Iterative Heuristic is based on a Greedy Heuristic. Therefore we can find properties of the resulting solution by exploiting the way the Greedy Heuristic works. First we note that the gross profit function for a product depends on demand, margin and operational constraints. Demand level and per-unit margin affect the maximum gross profit a product can generate if sufficient inventory is held. Operational constraints, such as case-pack sizes and delivery leadtime affect the curvature of the gross profit function. For example, a product with a smaller case-pack (batch size) has a higher slope of the gross profit curve for low inventory levels, and therefore can achieve the maximum gross profit with less inventory. These observations lead to the following theorems taken from Kök and Fisher (2007).

Products A and B belong to a subcategory with substitution rate $\delta \geq 0$. They are nonperishable. They are subject to the replenishment system described at the beginning of this subsection. The leadtime is zero. Demand for both products follow the same family of probability distributions. Effective demand for product A (B) has a mean D_A (D_B) and coefficient of variation ρ_A (ρ_B). Unless otherwise stated, $d_A = d_B, \rho_A = \rho_B, r_A = r_B, c_A = c_B$, and $b_A = b_B = 1$.

Theorem 3 *Consider products A and B. Let* $\tilde{\mathbf{f}}$ *denote the vector of facing allocations for all products in the subcategory other than A and B. If exactly one of the following conditions is met,*

(i) *All else is equal and* $d_{\mathrm{A}} > d_{\mathrm{B}}$*. The demand distribution is one of Poisson, Exponential or Normal distribution.*
(ii) *All else is equal and* $r_A - c_A \geq r_B - c_B$*.*
(iii) $w_{\mathrm{A}} \leq w_{\mathrm{B}}$,

then $f_{\mathrm{A}} \geq f_{\mathrm{B}}$ *in the final solution of the Iterative Heuristic.*

The implications of the first part of this theorem is clear: an allocation algorithm based on demand rates should work fairly well when products are differentiated by demand rates only. This is similar to the property of optimal assortments in the unconstrained problem in van Ryzin and Mahajan (1999). However, the above theorem proves additional results that the product with higher margin, or lower space requirement should be given priority in the assortment.

Theorem 4 *Consider products A and B. Let* $\tilde{\mathbf{f}}$ *denote the vector of facing allocations for all products in the subcategory other than A and B. If exactly one of the following conditions is met,*

(i) *All else is equal and* $\rho_{\mathrm{A}} < \rho_{\mathrm{B}}$,
(ii) *All else is equal,* $b_{\mathrm{A}} \geq 1$*, and* b_{B} *is an integer multiple of* b_{A},

then, the following holds. In the final solution of the Iterative Heuristic, if product B is included in the assortment then so is A (i.e.,$f_B > 0 \Rightarrow f_A > 0$).

Theorem 4 characterizes the impact of the operational characteristics of a product on the assortment choice. When one of the conditions of the Theorem 4 holds, i.e., when *B* has either a larger batch size or higher demand variability, due to limited shelf space, if *A* is not included in the assortment, neither is *B*. Since the maximum value of G_A is higher and the slope is higher for low inventory levels, the profit impact of first facing is higher for *A*, resulting in a higher rank in the ordered input list to the Greedy Heuristic. However, if both products are in the assortment, it is possible to have $f_B > f_A$ in the solution. The reason for this is that G_A reaches its maximum level quickly with the early facing allocations, whereas it takes more facings for *B* to reach its maximum. In such cases, allocation heuristics based on demand rates perform poorly. A reasonable rule of thumb based on these observations would be the following. First high demand rate products shall be included in the assortment, then more facings shall be allocated to the products that have more restrictive operational constraints.

We applied our estimation methodology (to be described in Sect. 5.2.2) and optimization methodology to the data from 37 stores and two categories. The categories include 34 subcategories or 234 SKUs. (AP) is solved for each category for a given category shelf space. The facing allocations for SKUs also determine the space allocation between subcategories. We compare the category gross profit of the recommended assortments with that of the current assortments at Albert Heijn.

The gross profits of the recommended system is 13.8 % higher than that of the current assortment. The financial impact of our methodology is a 52 % increase in pretax profits of Albert Heijn.

Other work on assortment planning with exogenous demand include Rajaram (2001). He develops a heuristic based on Lagrangian relaxation for the single period assortment planning problem in fashion retailing without consideration of substitution between products.

4.3 Assortment Planning Under Locational Choice

Gaur and Honhon (2006) study the assortment planning model under the locational choice demand model. The products in the category differ by a single characteristic that does not affect quality or price such as yogurt with different amounts of fat-content. The assortment carried by the retailer is represented by a vector of product specifications $(b_1, .., b_s)$ where s is the assortment size and $b_j \in [0, 1]$ denotes the location of product j. Each consumer is characterized by an ideal point in [0,1] and chooses the product that is closest to him or her. The coverage interval of product j is defined as the subinterval that contains the most preferred good of all consumers for whom the product yields a nonnegative utility. The first choice interval of product j is defined as the subinterval that contains the most preferred goods of all consumers who choose j as a first choice. To extend Lancaster's model to stochastic demand, the authors assume that customers arrive to the store according to a Poisson process and that the ideal points of consumers are independent and identically distributed with a continuous probability distribution on finite support [0,1]. Only unimodal distributions are considered, implying that there exists a unique most popular product, and that the density of consumers decreases as we move away from the most popular product.

The operational aspects of the problem are similar to the van Ryzin and Mahajan model reviewed in Sect. 4.1: all products are assumed to have identical costs and selling prices, there is a single selling period, inventory costs are derived from a newsvendor model: excess demand at the end of the period is lost and excess inventory is salvaged. The only difference is that there is a fixed cost associated with including a product in the assortment. This model is closely related to the marketing product line design models in the marketing literature and operations-marketing papers such as de Groote (1994).

Under static substitution (assortment-based substitution), a consumer chooses a first choice product given the assortment but without observing inventory levels and does not make a second choice if the first choice is not available. Under dynamic substitution, the consumer chooses a product (if any) among the available products. This is equivalent to choosing a first choice product from the assortment and then looking for the next best alternative (if any) if the first choice is not available. This is equivalent to stock-out based substitution with repeated attempts.

The paper characterizes the properties of the optimal solution under static substitution and develops approximations under dynamic substitution. We skip the details of the analysis and briefly discuss the results from this paper. The authors show that, under static substitution, the distance between products in the optimal assortment are large enough so that there is no substitution between them. The most popular product, the one that would be located at the mode of the distribution is not included in the assortment when the economies of scale enjoyed by the most popular product is overcome by the diseconomies of scale it created for the other products. This property contrasts with the property of the optimal assortments under the MNL model (Theorem 2). We believe that the difference is not because of the different choice model, but because the problem considered here is a product line design problem at its heart. The authors find that the retailer may choose not to cover the entire market due to fixed costs. An analogous result is obtained under the MNL model as well, but that is purely due to economies of scale created for more popular products by not including some products in the assortment. Whereas in this model, it is optimal to cover the entire market when fixed costs are not present.

The problem is more complex under the dynamic substitution problem, as it is under other demand models. The profits computed under the static substitution assumption provides a lower bound to the dynamic problem, since it does not capture the profits from repeated attempts of the stock-out based substitution. An upper bound is obtained by solving a relaxation of the problem. Namely, the retailer gets to observe the ideal points of all arriving customers before allocating inventory to customers to maximize the profits. This is similar to Bassok et al. (1999) where consumers do not directly choose a product, but they are assigned a product (if any) either according to an exogenous rule or the retailer's decisions. Clearly, the retailer can generate more profits by doing the allocation itself rather than following the choices of the customers arriving in a random process. The solutions to these bounds are also proposed as heuristic approaches. In a numerical study, the authors make the following observations. Both heuristics generate solutions that are 1.5 % within the optimal solution on average. This suggests that the static substitution solution, which is easier to obtain, would serve as a good approximation in most cases. Dynamic substitution has the greatest impact when demand is low, customer distribution in the attribute space is heterogenous, and consumers are willing to substitute more. The retailer provides higher variety under dynamic substitution than under static substitution and locates products closer to each other so that a consumer can derive positive utility from more than one product. The firm offers more acceptable alternatives to the customers whose ideal product is located in areas where consumer density is high.

There are other papers that formulate mathematical models for selecting optimal assortments when customer heterogeneity is represented by locational choice. McBride and Zufryden (1988) deal with manufacturer's product line selection which require specification of product attributes and Kohli and Sukumar (1990) deal with the retailer's problem of choosing an assortment from a set of products.

Alptekinoğlu et al. (2012) extend the Hotelling-Lancaster locational choice model for studying the assortment planning problem for a category of horizontally differentiated products. They assume that consumer preferences are distributed along a straight line, and the disutility costs due to substitution are asymmetric and convex with respect to distance. They show that when preferences follow a unimodal distribution, the prices and market share of the products drop with distance in respect to the product that covers the mode (or the most popular product). They show that their approach leads to exact solutions when consumer tastes are distributed discretely. For continuous distributions, they propose a shortest path formulation, which can be computed efficiently.

4.4 Assortment Planning in Decentralized Supply Chains

The assortment planning papers reviewed until this section are single location models. There has been some recent work exploring assortment planning issues in two-tier supply chains. Aydin and Hausman (2003) consider the assortment planning problem with MNL (i.e. the van Ryzin and Mahajan model) in a decentralized supply chain with one supplier and one retailer. They find that the retailer chooses a narrower assortment than the supply chain optimal assortment since her profit margins are lower than that of the centralized (vertically integrated) supply chain. The manufacturer can induce coordination by paying the retailer a per-product fee, resembling the slotting fees in the grocery industry, while making both parties more profitable.

Singh et al. (2005) study the effect of product variety on supply chain structures, building on the van Ryzin and Mahajan model. In the traditional channel, the retailers stock and own the inventory, whereas in the drop-shipping channel, the wholesaler stocks and owns the inventory and ships the products directly to customers after the customers place an order at a retailer. Drop-shipping is a common practice in internet retailing: it offers the benefits of risk pooling when there are multiple retailers, but retailers have to pay a per unit fee for drop-shipping. As a result, product variety in the drop-shipping channel is higher than the traditional channel when drop-shipping fees are low and number of retailers is large. The authors derive conditions on the parameters under which the retailers or the wholesaler or both prefer the drop-shipping channel. They also study a vertically integrated firm with multiple retailers and find that a hybrid supply chain structure may be optimal for some parameter combinations: the popular products are stocked at the retailer while the less popular products are stocked at the warehouse and drop-shipped to the customers. The assortment size at the retailer gets smaller as the number of retailers increase or the drop-shipping costs decrease.

Kurtulus and Toktay (2007) compare the traditional category management and category captainship in a setting with two products and deterministic demand under a shelf space constraint. In category captainship, one of the vendors is assigned as the category captain and the pricing and assortment decisions are delegated to her.

The argument for category captainship is that the leading manufacturer in a category may have more experience with the category and resources than the retailer. They find that the assortment may be narrower under category captainship, because the noncaptain brand may be priced out of the assortment. Kurtulus (2005) considers the impact of category captainship under three types of contracts in a setting similar to the van Ryzin and Mahajan model. While the resulting assortment is still in the popular assortment set under the target profit and target sales contracts, it is in the least popular assortment set under the target variety contract.

4.5 Dynamic Assortment Planning

All of the assortment planning papers reviewed in the previous sections consider static assortment planning problems and do not consider revising or changing assortment selection as time elapses. This makes sense for fashion and apparel retailers, because long development, procurement and production lead times constrain retailers to make assortment decisions in advance of the selling season. With limited ability to revise product assortments, academics and industry practitioners focused on optimizing the production quantities in order to delay the production of those products that have high demand uncertainty (e.g., Fisher and Raman 1996). However, innovative firms such as Zara (Spain), Mango (Spain), and World Co. (Japan) created highly responsive and flexible supply chains and cut the design-to-shelf lead time down to 2–5 weeks, as opposed to 6–9 months for a traditional retailer, which enabled them to make design and assortment selection decisions during the selling season. Raman et al. (2001) describes how such short response times are achieved at World Co. through process and organizational changes in the supply chain. Learning the fashion trends and responding with an updated product selection is most critical for these high fashion companies.

Allowing changes in the assortment during a single selling season introduces several new issues. The products put in the store this week can't be removed next week and hence condition the decisions this week; there may be costs associated with adding new products or dropping products from the assortment; it may be optimal to put products in the stores to learn about the demand, even if it isn't optimal to do so given the current knowledge.

Caro and Gallien (2007) formulate the dynamic assortment problem faced by these retailers: At the beginning of each period, the retailer decides which assortment should be offered and gathers demand data for the products carried in the assortment in each period. There is a budget constraint that limits the number of products offered in each period to K. Due to design-to-shelf lead time, an assortment decision can be implemented only after l periods. This problem relates to the classical exploration versus exploitation trade-off. The firm must decide whether to optimize revenues based on the current information (exploitation), or try to learn more about the demand of products not in the assortment with the hope of identifying popular products (exploration).

The authors make several assumptions for tractability. The demand for a product is independent of the demand or the availability of the other products (i.e., there is no substitution between products or correlation in demand). The demand rate for each product is constant throughout the season. There is a perfect inventory replenishment process, therefore there are no lost sales or economies of scale in the operating costs. More importantly, no products carry over from period to period, therefore it is feasible to change the assortment independent of the previous assortment. There are no switching costs. Some of these assumptions are relaxed later.

The demand for product $j \in N$ is from a stationary Poisson process throughout the season. The rate of arrival λ_j is unknown and actual demand is observed only when the product is included in the assortment. The retailer uses a Bayesian learning mechanism: he starts each period with a prior belief that λ_j is distributed according to a Gamma distribution with shape parameter m_j and scale parameter α_j. Suppose that product j is included in the assortment and observed demand is d_j. The prior distribution of λ_j is updated as $Gamma(m_j + d_j, \alpha_j + 1)$. The mean of this distribution is the average sales of product j throughout the periods it is carried. Let $\mathbf{f} = (f_1, \ldots, f_n)$ be a vector of binary variables indicating whether the product is in the assortment and F the set of feasible assortments, $F = \{\mathbf{f} : \sum_{j \in N} f_j \leq K\}$. Similarly, let $\mathbf{m}, \boldsymbol{\alpha}$, and $\mathbf{d}$ denote the vectors of m_j, α_j, d_j, respectively. Assume that assortment implementation leadtime l is zero.

The dynamic programming formulation is

$$J_t^*(\mathbf{m}, \alpha) = \max_{\mathbf{f} \in F} \sum_{j \in N} f_j r_j E[\lambda_j] + EJ_{t+1}^*(\mathbf{m} + \mathbf{d} \cdot \mathbf{f}, \alpha + \mathbf{f}).$$

Since solution of this dynamic program can be computationally overwhelming, the authors propose a Lagrangian relaxation (of the constraint on the number of products in the assortment) and the decomposition of weakly coupled dynamic programs to develop an upper bound. Performance of two heuristics are compared. The index policy balances exploration by including high expected profit products and exploitation by including products with high demand variance in a single-period look ahead policy. The greedy heuristic selects in each period the K products with the highest expected profits. The index policy is near optimal when there is some prior data on demand available and outperforms the greedy heuristic especially with little prior information about demand or the leadtime. The paper then demonstrates that the heuristics perform well when there are assortment switching costs, demand substitution, and a positive implementation lag.

Another learning method that Zara and other high-fashion companies employ is learning the attributes of the high selling products. That is, if a certain color is hot this season, and products with a special fabric are selling relatively well, the prior distribution of the demand for a product with that fabric-color combination can be updated, even if the product were never included in the assortment before. The attribute-based estimation method by Fader and Hardie (1996) mentioned in

Sect. 5.1 can be instrumental in estimating the demand for new products in this setting.

Rusmevichientong et al. (2010) develop algorithms to compute the optimal assortment under multinomial logit demand and capacity constraints. They derive structural insights on the optimal assortment for the static case, and utilize it to develop an adaptive policy for the dynamic problem, where the algorithm learns demand parameters form past data and chooses the optimal assortment based on that. They find that their algorithm performs well on being applied to sales data from an online retailer.

Ulu et al. (2012) study the dynamic assortment problem under horizontal differentiation, when consumer preferences are distributed according to the locational choice model. They assume that the firm knows where customers are located, but is unaware of their probability distribution. They model the problem using a discrete-time dynamic program, where in each period the retailer chooses an assortment and set of prices to maximize expected profits over the entire horizon, and customers choose the utility maximizing product from the assortment. The retailer updates beliefs on the distribution of customers in a Bayesian fashion. Under this scenario, they show that it is possible to partially order assortments based on their information content. They demonstrate that it might be optimal for the retailer to alternate between exploration and exploitation, and sometimes offer sub optimal loss producing assortments in a bid to learn valuable information about consumer preferences.

Bernstein et al. (2011) present a novel model exploring dynamic assortment decisions in a setting with multiple heterogeneous customer segments. They show that rationing products to some customer segments may be optimal. This insight is different from those obtained in the revenue management literature, as the rationing outcome is not due to differences in costs or prices, but due to the interplay between heterogeneity in customer segments and limited inventories. They demonstrate the potential impact of assortment customization based on a real data set obtained from a large fashion retailer. They find that the revenue impact of assortment customization can be significant indicating its potential as another lever for revenue maximization in addition to pricing.

Saure and Zeevi (2013) consider the interesting case where a retailer tries to learn about consumer preferences by strategically offering different assortments. The main tradeoff facing the retailer is to balance the value of learning with the goal of maximizing revenues. They study a family of stylized assortment planning problems under this scenario, and develop a family of policies that balance this tradeoff. Their major finding is that the optimal policy limits experimentation with suboptimal products, thereby reducing the impact of experimentation on revenues.

4.6 Competitive Assortment Models

Cachon and Kök (2007) study the assortment planning problem with multiple merchandise categories and basket shopping consumers (i.e., consumers who desire to purchase from multiple categories). They present a duopoly model in which retailers choose prices and variety level in each category and consumers make their store choice between retail stores and a no-purchase alternative based on their utilities from each category. The common practice of category management (CM) is an example of a decentralized regime for controlling assortment because each category manager is responsible for maximizing his or her assigned category's profit. Alternatively, a retailer can make category decisions across the store with a centralized regime. They show that CM never finds the optimal solution and provides both less variety and higher prices than optimal. In a numerical study, they demonstrate that profit loss due to CM can be significant. Finally, they propose a decentralized regime that uses basket profits, a new metric, rather than accounting profits. Basket profits are easily evaluated using point-of-sale data, and the proposed method produces near-optimal solutions.

Hopp and Xu (2008) consider a static approximation of the assortment planning problem under stock-out substitution. They model demand using fluid networks and obtain a mapping between service and inventory, which allows them to analyze the previously intractable, joint assortment, inventory and pricing problem in both competitive and non-competitive scenarios. They show that the static approximation models the dynamic scenario very closely, and obtain several interesting structural insights under duopolistic competition. First, they find that under joint price and inventory competition, prices are lower, while demand and inventory levels are higher. Second, they observe that under joint price and assortment competition, prices and variety offered by each retailer are both lower. However, the total number of products and the aggregate inventory levels in a duopoly market and both higher than in a monopolistic market.

Kök and Xu (2011) study assortment planning and pricing for a product category with heterogeneous product types from two brands. They model consumer choice using the Nested Multinomial Logit framework with two different hierarchical structures: a brand-primary model in which consumers choose a brand first, then a product type in the chosen brand, and a type-primary model in which consumers choose a product type first, then a brand within that product type. They find that optimal (centralized) and competitive (decentralized between brands) assortments and prices have quite distinctive properties across different models. Specifically, with the brand-primary model, both the optimal and the competitive assortments for each brand consist of the most popular product types from the brand. They extend the structural properties of assortment decisions characterized by van Ryzin and Mahajan (1999) to the case of Nested Logit. Under the brand primary model, structure remains the same under competitive and centralized regimes. The type-primary choice model, however, leads to a structural difference: The optimal and the competitive assortments for each brand may not always consist of the most popular product types of the brand. Instead, the overall assortment in the category consists of a set of most popular product types. Further, due to the combinatorial nature of the type-primary model, the existence of equilibrium may not be guaranteed. This paper also characterizes the optimal pricing of products. They find that a lower price should be charged for more popular product types due to economies of scale. Under competition, the brand with the higher market share would charge higher prices.

Besbes and Saure (2011) study the assortment problem under a duopoly, when consumers make their purchase decisions with full knowledge of the retailers' assortments. They show that when prices are exogenous, and the products carried by the retailers are exclusive, the number of equilibria are bounded, and the retailers always prefer the same equilibrium. When the assortments overlap, they show that an equilibrium may or may not exist, and the number of equilibria might increase exponentially with the number of products. Under the scenario of joint assortment and price competition, they show that at most one equilibrium exists. Finally, they demonstrate that competition leads to lower prices and expanded variety, as compared to a monopolistic setting.

Martínez-de-Albéniz and Roels (2011) consider shelf-space competition in a multi-supplier retail outlet. They find that when retailers allocate shelf space between products based on sales velocity and margins, and suppliers set wholesale prices to maximize the shelf space they are allocated, they tend to keep margins high. Moreover, the incentives of the two parties are misaligned, leading to suboptimal prices and shelf space allocations. Additionally, they find that the impact of suboptimal pricing far outweighs the effect of suboptimal shelf space allocation.

Kök and Martínez-de-Albéniz (2013) study the impact of quick response capabilities of supply chains on product variety in a competitive environment. In industries where customer needs quickly change, retailers such as Zara can postpone their assortment decisions (amount of variety, balance across categories) to close-to-season or in-season due to shorter design-to-shelf lead times. The authors study how assortment competition depends on the postponement capabilities of retailers. They develop a stylized model where two retailers choose their assortment breadth either before or after market characteristics are revealed. They find that slower retailers provide a higher variety and being fast is equivalent to offering 30–50 % more variety.

4.7 Assortment Planning Models with Multiple Categories/Stores

Although research has primarily focused on single category choice decisions, there is recent research that examines multiple category purchases in a single shopping occasion by modeling the dependency across multi-category items explicitly (see

Russell et al. 1997 for a review). Manchanda et al. (1999) find that two categories may co-occur in a consumer basket either due to their complementary nature (e.g., cake mix and frosting) or due to coincidence (e.g., similar purchase cycles or other unobserved factors). Bell and Lattin (1998) show that consumers make their store choice based on the total basket utility. Fixed costs for each store visit (e.g., search and travel costs) provide an intuitive explanation for why consumers basket shop. Bell et al. (1998) use market basket data to analyze consumer store choices and explicitly consider the roles of fixed and variable costs of shopping.

Baumol and Ide (1956) study the notion of right level of variety in a very stylized model. The retailer chooses N, the number of different product categories to offer. Consumer utility is increasing in variety, but decreasing in in-store search costs (which increases with N). Therefore for each consumer there is a range of N that makes the store attractive for shopping. The operating cost is the sum of inventory costs per category from an EOQ model and handling costs that is concave increasing in N. The resulting retailer profit function is not well-behaved, therefore profit maximizing level of variety is difficult to characterize and the insights from this model are fairly limited.

There are two papers that consider assortment planning with multiple categories in more detail. Agrawal and Smith (2003) extend the Smith and Agrawal (2000) model and the analysis described in Sect. 4.2.1 to the case where customers demand sets of products. Cachon and Kök (2007) compare the prices and variety levels in multiple categories under category management to the optimal variety levels in the presence of basket shopping consumers.

The modeling and solution approach in Agrawal and Smith (2003) is very similar to their earlier work. Each arriving customer demands a purchase set. If the initially preferred purchase set is not available, the customer may do one of the following: (a) substitute a smaller set that does not contain the missing item, (b) substitute a completely different purchase set, (c) not purchase anything. This behavior is governed by substitution probability matrices. The demand for each set considering the substitution demand from other sets is characterized as in Eq. (8.4). The profit maximization problem is formulated as a mathematical program. For a customer to purchase any set, all the items in the set have to be available. Therefore, the expected profit is much more sensitive to percentage of customers who purchase in sets, the average size of a purchase set, and the substitution structure and parameters. The following observations from numerical examples are quite interesting.

Profits under adjacent substitution structure is much higher than that under random substitution, because under adjacent substitution stocking every other set in the list would result in lower lost sales than that under random substitution. As the percentage of customers who purchases in sets increases (while keeping the total demand constant), the optimal assortment size increases (decreases) if the fixed cost of including a product is low (high). Profits increase with substitution rate δ. Finally, optimizing the category by disregarding the substitution and the purchase sets can result in considerably lower profits than optimal.

Cachon and Kök (2007) work with a stylized model to develop managerial insights regarding the assortment planning process in an environment with multiple categories. Consider two retailers X and Y that carry two categories of goods. Retailer r offers n_{rj} products and sets its margin p_{rj} in category j. The consumer choice model is based on a nested Multinomial Logit (MNL) framework. A consumer's utility from purchasing product i in category j at retailer r is $u_{rji} = v_{rji} - p_{rj} + \varepsilon$ where v_{rji} is the expected utility from the product less the unit cost of the product and ε is i.i.d with Gumbel distribution with zero mean. There are three types of consumers in the market that are characterized by the contents of their shopping baskets: type 1 consumers would like to buy a product in category 1 only, type 2 consumers would like to buy a product in category 2 only, type b consumers are basket shoppers and would like to buy a product from both categories. Consumers buy exactly one unit of one product in every category included in their basket.

The authors show that the choice probability of a non-basket shopper between retailers X, Y and a no-purchase alternative can be written using the nested MNL model as follows:

$$s_{rj} = \frac{A_{rj}}{A_{xj} + A_{yj} + Z_j} \quad \text{for} \quad r = x, y, \quad \text{and} \quad j = 1, 2,$$

where A_{rj} is the attractiveness function for each alternative (an aggregate function of price and variety level). Using the nested MNL results of Ben-Akiva and Lerman (1985), as described in Sect. 3.2, it can be expressed as

$$A_{rj} = e^{-p_{rj}} \sum_{i=1}^{n_{rj}} e^{v_{rji}}, \quad \text{for} \quad r = x, y.$$

Now, consider a basket-shopping consumer. A basket-shopping consumer chooses retailer r only if she prefers the assortment at r for both categories. As a result, the probability that a basket shopper chooses retailer r is

$$s_{rb} = s_{r1} s_{r2} \quad \text{for} \quad r = x, y. \tag{8.7}$$

This is a multiplicative basket-shopping model, as a retailer's share of basket shoppers is multiplicative in its share in each category. An additive model for this problem has been discussed in Kök (2003).

The common practice of category management (CM) is an example of a decentralized regime for controlling assortment because each category manager is charged with maximizing profit for his or her assigned category. Since basket shoppers' store choice decision depends on the prices and variety levels of other categories, one category's optimal decisions depends on the decisions of the other categories. Hence, a game theoretic situation arises. CM can be interpreted as an explicit non-cooperative game between the category managers, since each category manager is responsible exclusively for the profits of her own category.

Alternatively, it can be interpreted as an iterative application of single category planning where each category's variety level is optimized assuming all other assortment decisions for the retailer are fixed. Decentralized regimes such as CM are analytically manageable but they ignore (in their pure form) the impact of cross-category interactions. Centralized regimes account for these effects but it is extremely difficult, in practice, to design a model to account for all cross-category effects, to estimate its parameters with available data and solve it.

The authors show that if there are any basket shoppers, CM provides less variety and higher prices than centralized store management. CM can lead to poor decisions because the category manager does not sufficiently account for how his or her decisions influences total store traffic. These results hold both for a single retailer and in duopoly competition. Numerical examples demonstrate that the profit loss due to CM can be significant. The dominant strategy for each retailer is to switch to centralized management.

To address the potential problem with a decentralized approach to assortment planning, we propose a simple heuristic that retains decentralized decision making (category managers optimize their own categories' profit) but adjusts how profits are measured. To be specific, instead of using an accounting measure of a category's profit, the authors define a new measure called *basket profits*. Basket profits can be estimated using point-of-sale data. It enables CM to approximately measure the true marginal benefits of merchandising decisions and lead to near-optimal profits. This analytical approach is an attractive alternative relative to ad-hoc coordination across category managers.

Fisher and Vaidyanathan (2014), consider the assortment localization problem, of choosing assortments that can vary by store, subject to a maximum number of different assortments. They model a SKU as a set of attributes and also model possible substitutions when a customer's first choice is not in the assortment. estimate demand and substitution probabilities from sales history using maximum likelihood estimation. They apply maximum likelihood estimation to sales history of the SKUs currently carried by the retailer to estimate the demand for attribute levels and substitution probabilities, and from this, the demand for any potential SKU, including those not currently carried by the retailer. They develop several heuristics for choosing SKUs to be carried in an assortment, and apply this approach to optimize assortments for three real examples: snack cakes, tires and automotive appearance chemicals. A portion of their recommendations for tires and appearance chemicals were implemented and produced sales increases of 5.8 % and 3.6 % respectively, which are significant improvements relative to typical retailer annual comparable store revenue increases.

5 Demand Estimation

In this section, we briefly discuss the estimation of the demand models specified in Sect. 3. The estimation method depends on the type of data that is available.

5.1 Estimation of the MNL

5.1.1 With Panel Data

Starting with the seminal work of Guadagni and Little (1983), an enormous number of marketing papers estimated the parameters of the MNL model to understand the impact of marketing mix variables on demand. These papers use panel data in which the purchasing behavior of households over time are tracked by the use of store loyalty cards. Consider the purchase decision of the household that visited the store in time t. The systematic component of the utility u_{jt} is specified as a linear function of m independent variables including product specific intercepts, price, an attribute of product j, loyalty of the household to the brand of product j (measured as exponentially weighted average of binary variables indicating whether or not the household purchased this brand). Let $x_{jt} = (x_{jt1}, x_{jt2}, .., x_{jtm})$ denote the vector of these attributes for the household's shopping trip at time t, S_t denote the assortment at time t including the no-purchase option, and $\beta = (\beta_1, .., \beta_m)$ denote the vector of common coefficients.

$$u_{jt} = \beta^T x_{jt}, \qquad j = 0, 1, .., n.$$

The outcome of the choice experiment by a household in time t is

$$y_{jt} = \begin{cases} 1, & \text{if product } j \text{ is chosen in time } t \\ 0, & \text{otherwise} \end{cases}$$

Given u_{jt} it is possible to compute the choice probabilities according to MNL formula (8.1). To obtain the maximum likelihood estimates (MLE) for the coefficients, we can write log of the likelihood function by multiplying the probability of observing the choice outcome across all t:

$$L(\beta) = \sum_t \sum_j y_{jt} \left(\beta^T x_{jt} - \ln \sum_{k \in S_t} e^{\beta^T x_{kt}} \right).$$

McFadden (1974) shows that the log-likelihood function is concave, therefore any nonlinear optimization technique can be used to find the MLE estimate of β. Fader and Hardie (1996) suggest the use of more of the product's attributes and dropping product-specific dummy variables in x_j in the estimation. They argue that this results in a more parsimonious estimation method as the number of coefficients to be estimated would not grow with number of products but with number of significant characteristics. Moreover, this approach enables estimation of the demand for new products.

Extensions of this model such as Chiang (1991), Bucklin and Gupta (1992), and Chintagunta (1993) also investigate whether to buy, and how much to buy decisions of households. In these papers, the whether-to-buy decision is modeled as a binary choice between the no-purchase alternative and the resulting utility from the product choice and quantity decisions in a nested way. Chong et al. (2001) extend the classical Guadagni and Little (1983) model using a nested MNL model, including three new brand-width measures that capture the similarities and the differences among products within and across brands.

Multiplicative Competitive Interactions (MCI) model offers a viable alternative to MNL. Although less popular than MNL, it is used in the marketing area to study market share games (e.g. Gruca and Sudharshan 1991) and it has empirical support. See Cooper and Nakanishi (1988) for a detailed discussion and estimation methods.

5.1.2 With Sales Transaction Data

Consider the demand process in the van Ryzin and Mahajan model, where consumer arrivals follow a Poisson process with rate λ and consumers select an alternative based on the MNL model. Our goal is to estimate λ and β from sales data. Sales transactions are the records of the purchasing time and the product choice for each customer who made a purchase. This is an incomplete data set in the sense that only the arrivals of customers who made a purchase are recorded. Define a period as a very small time interval such that the probability of having more than one customer arrival in a period is zero. Let t denote the index of periods. There is a sales record for a period only if a purchase is made in that period. It is impossible to distinguish a period without an arrival, from a period in which there was an arrival but the customer did not purchase anything. Therefore, the approach described above cannot be used.

The *Expectation-Maximization* (EM) algorithm is the most widely used method to correct for missing data. Proposed by Dempster et al. (1977), the EM method uses the complete-likelihood function in an iterative algorithm. Talluri and van Ryzin (2004) describe an estimation approach based on this method in the context of airline revenue management, but the algorithm is applicable to the retail setting described in Sect. 4.1. Let P denote the set of periods that there has not been a purchase made and $a_t = 1$ if there has been a customer arrival in period t. The unknown data is $(a_t)_{t \in P}$. We start with arbitrary (λ, β). The E-step replaces the incomplete data with their estimates. That is, we find the expectation of a_t for all $t \in P$ given the current estimates (λ, β). The M-step maximizes the complete-data likelihood function to obtain new estimates. The likelihood function is similar to that in the previous subsection, but includes the arrival probabilities λ. The procedure is repeated until the parameter estimates converge. Greene (1997) shows that the procedure converges under fairly weak conditions. If the expected log-likelihood function is continuous in the parameters, Wu (1983) shows that the limiting value of the procedure would be a stationary point of the incomplete-data log-likelihood function. The advantage of the procedure is that maximizing the complete-data likelihood function is much easier than maximizing an incomplete-data likelihood function.

Musalem et al. (2010) use store-level data and partial information on product availability to estimate consumer demand under stock-out based substitution. They develop a structural demand model that simulates the effect of stock-outs using a time-varying set of available alternatives, and is able to capture very flexible substitution patterns. They demonstrate how their model can be used to quantify lost sales and provide insights on the financial consequences of stock-outs. Finally, they suggest how price promotions can be used effectively to counter some of the negative economic impact.

Vulcano et al. (2012) focus on the problem of estimating demand model when only sales transaction data are available. They model demand by combining a poisson arrival process with a multinomial choice process. Instead of estimating the arrival and choice parameters simultaneously by maximizing an intractable likelihood function, they treat observed demand as incomplete realizations of primary demand, and utilize an Expectation-Maximization approach to develop simple and efficient algorithms to estimate the model parameters. They test the utility of their approach on one simulated and two industry data sets.

Jain et al. (2014) consider how sales transaction timing data can lead to better demand estimates. They find that the optimal order quantity is higher when the retailer takes into account actual stock-out times, as compared to the case where demand is fully observed. However, in most cases, where the demand uncertainty is high, and the margins are low, the extent of over-ordering with timing data tends to be lower than that with only stock-out event data. They demonstrate using numerical simulations, that the use of stock-out timing data reduces the loss in expected profits by 74.8 % as compared to the case where only stock-out events are observed.

5.1.3 With Sales Summary Data

The information available in sales data is different from the panel data in several ways, hence requires a different approach. One possibility is the approach in Kök and Fisher (2007), which will be described here. The data typically available for estimating the parameters of a demand model includes the number of customers visiting each store on a given day, sales for each product-store-day, as well as the values of variables that influence demand such as weather, holidays, and marketing variables like price and promotion. At Albert Heijn, the data set included SKU-day-store level sales data through a period of 20 weeks for seven merchandise categories from 37 Albert Heijn stores. For each store-day, the number of customers visiting the store is recorded. For each SKU-day-store, sales data comprised of the number of units sold, the number of customers that bought that product, selling price, and whether the product is on promotion or not. In addition, we have daily weather data and a calendar of holidays (e.g., Christmas week, Easter, etc.). The categories are cereals, bread spreads, butter and margarine, canned fruits, canned vegetables, cookies, and banquet sweets. There were 114 subcategories in these seven categories. The size of subcategories varies from 1 to 29 SKUs, with an average of 7.7 and a standard deviation of 5.7.

The model of consumer purchase behavior is based on three decisions: (1) whether or not to buy from a subcategory (*purchase-incidence*), (2) which variant to buy (*choice*) given purchase incidence, and (3) how many units to buy (*quantity*).[1] This hierarchical model is quite standard in the marketing literature and commonly used with panel data.

The demand for product j is

$$D_j = K(PQ)_j = K\pi p_j q_j \tag{8.8}$$

where K is the number of customers that visit the store at a given day, $(PQ)_j$ is the average demand for product j per customer, π is the probability of purchase incidence (i.e., the probability that a customer visiting the store buys anything from the subcategory of interest), p_j is the choice probability (i.e., the probability that variant j is chosen by a customer given purchase incidence), and q_j is the average quantity of units that a customer buys given purchase incidence and choice of product j.

The purchase incidence is modeled as a binary choice:

$$\pi = \frac{e^v}{1 + e^v} \tag{8.9}$$

where v is the expected utility from the subcategory that depends on the demand drivers in the subcategory.

The product choice is modeled with the Multinomial Logit framework, where p_j are given by (8.1). The average utility of product j to a customer, u_j, is assumed to be a function of product characteristics, marketing and environmental variables.

Let subscript h denote store index, and t denote time index (i.e., day of the observation).

We compute p_{jht} from the sales data as the ratio of number of customers that bought product j to number of the customers that bought any product in the subcategory at store h on day t. At Albert Heijn, price and promotion are the variables influencing u_j. We fit an ordinary linear regression to the log-centered transformation of (8.1) (see Cooper and Nakanishi 1988 for details) to estimate $\delta_j^C, \alpha_1^C, \alpha_2^C$, and $\theta_k^C, k = 1, .., n$.

$$\ln\left(\frac{p_{jht}}{\overline{p}_{ht}}\right) = u_j = \delta_j^C + \sum_{k \in N} \theta_k^C I_{jk} + \alpha_1^C \left(R_{jht} - \overline{R}_{ht}\right) + \alpha_2^C \left(A_{jht} - \overline{A}_{ht}\right), \quad \text{for all } j \in S \tag{8.10}$$

[1] This hierarchical model of choice is similar to Bucklin and Gupta (1992) that models the first two decisions with an additional focus on the segmentation of customers and Chintagunta (1993) that models all three decisions. Both papers work with household panel data, whereas we work with daily sales data.

where $\overline{p}_{ht} = \left(\prod_{j\in S} p_{jht}\right)^{1/|S|}$, $I_{jk} = \{1, \text{if } j = k; 0 \text{ otherwise}\}$, R is price, $\overline{R}$ is average price in the subcategory, $A_{jht} = \{1, \text{if product } j \text{ is on promotion on day } t \text{ at store } h;$ $0, \text{otherwise}\}$, and $\overline{A}$ is average promotion level in the subcategory. It is straightforward to incorporate variables other than price and promotion into this approach.

We compute π_{ht}, the probability of purchase-incidence for the subcategory, from sales data as the ratio of number of customers who bought any product in S to the number of customers visited the store h on day t. We use the following logistic regression equation to estimate $\alpha_0^\pi, \alpha_1^\pi, \alpha_2^\pi, \alpha_{4t}^\pi, \gamma_k^\pi$, $k = 1, ..6$, and $\beta_l^\pi, l = 1, .., 14$ in (8.11).

$$\ln\left(\frac{\pi_{ht}}{1-\pi_{ht}}\right) = v = \alpha_0^\pi + \alpha_1^\pi T_t + \alpha_2^\pi HDI_t + \sum_{k=1}^{6} \gamma_k^\pi D_t^k + \alpha_{4t}^\pi \overline{A}_{ht} + \sum_{l=1}^{14} \beta_l^\pi E_t^l \quad (8.11)$$

where T is the weather temperature, HDI (Human Discomfort Index) is a combination of hours of sunshine and humidity, D^k are day of the week 0–1 dummies and E^l are holiday 0–1 dummies for Christmas, Easter, etc. Other variables could be used appropriately in a different context.

We compute q_{jht} from sales data as the number of units of product j sold divided by the number of customers who bought product j at store h on day t and use linear regression to estimate $\alpha_{0j}^Q, \alpha_{1j}^Q, \alpha_{2j}^Q$, and $\beta_{jl}^Q, l = 1, \ldots, 14$ in (8.12).

$$q_{jht} = \alpha_{0j}^Q + \alpha_{1j}^Q A_{jht} + \alpha_{2j}^Q HDI_t + \sum_{l=1}^{14} \beta_{jl}^Q E_t^l, \quad \text{for all } j \in S \qquad (8.12)$$

In the grocery industry, K_{ht}, the daily number of customers who made transactions in store h on day t is a good proxy for the number of customers who visited the store. We use log-linear regression to estimate $\alpha_{0h}^K, \alpha_{1h}^K, \alpha_{2h}^K, \gamma_k^K, k = 1, .., 6$, and β_{1l}^K, $l = 1, \ldots, 14$ in (8.13).

$$\ln(K_{ht}) = \alpha_{0h}^K + \alpha_{1h}^K T_t + \alpha_{2h}^K HDI_t + \sum_{k=1}^{6} \gamma_k^K D_t^k + \sum_{l=1}^{14} \beta_{1l}^K E_t^l \qquad (8.13)$$

This four stage model of demand estimation has been tested for quality of fit and prediction for multiple stores and subcategories. The average of mean absolute deviation (MAD) across all products, subcategories and stores is 67 % in the fit sample and 74 % in the test sample. Average bias of our approach is 0 % and −9 % in fit and test samples, respectively. The current method used at Albert Heijn is estimating $(PQ)_j$ for each SKU directly via logistic regression with similar explanatory variables. The MAD of this method is 72 % and 94 % and average bias is −43 % and −30 % in the fit and test samples, respectively.

5.2 Estimation of Substitution Rates in Exogenous Demand Models

5.2.1 Estimation of Stockout-Based Substitution

Anupindi et al. (1998) estimate the demand for two products and the substitution rates between them using data from vending machines. They assume that consumers arrive according to a Poisson process with rate λ and choose product A (B) as their first choice product with probability p_A (p_B) and substitute according to an asymmetric substitution matrix $\begin{bmatrix} 0 & \alpha_{AB} \\ \alpha_{BA} & 0 \end{bmatrix}$. The demand for product A when B is not available is Poisson with rate $\lambda(p_A + p_B\alpha_{BA})$.

They consider two information scenarios. In the first one, so-called perpetual inventory data, each sales transaction and the exact time that each product runs out of stock (if they do) is observed. In this case, it is not difficult to write down the log-likelihood function and maximize it to obtain the MLE estimates. They show that the timing of the stockouts and the sales volume before and after those times are sufficient statistics. Therefore, it is not necessary to trace each sales transaction. This result of course would not hold if the arrival process were a nonstationary process.

In the second information scenario, so-called periodic review data, the stockout times of the products are not observed, but whether or not they are in-stock at the time of replenishment is known. We encounter an incomplete data problem, and again we can use the EM algorithm briefly discussed in Sect. 5.1.2 to correct for the missing data (i.e., the stockout times). To be able to generalize the methodology to more than two products, it is necessary to make further assumptions. The authors restrict the substitution behavior to a single-attempt model, i.e., no repeated attempts are allowed and they estimate the parameters for a problem with six products. Their results show that naive demand estimation based on sales data is biased, even for items that rarely stockout. They also find significant differences in the substitution rates of the six brands.

Anupindi et al. (1998) estimate stationary demand rates (i.e., do not consider a choice process) and a substitution matrix. Talluri and van Ryzin (2004) estimate demand rate and the parameters of the MNL choice model (λ, β) but do not consider a substitution matrix. Kök and Fisher (2007) generalize these two approaches and propose a procedure that simultaneously estimates the parameters of the MNL model, on which the consumer's original choice is based, and a general substitution probability matrix.

5.2.2 Estimation of Assortment-Based Substitution

Some retailers do not track inventory data. Some others do, but there is empirical evidence that the inventory data may not be accurate (e.g. DeHoratius and Raman 2008). Hence, sales data may be the only source of information in some cases. Here we review the methodology proposed by Kök and Fisher (2007) to estimate substitution rates using sales data. We assume that substitution structure (i.e., the type of the matrix) is known, and we only need to estimate the substitution rate δ. We demonstrate the method for the proportional substitution matrix, that is assume α_{kj} is given by (8.2).

The methodology can be explained briefly as follows. Suppose that a store carries assortment $S \subset N$ with 100 % service rate (i.e., no stockout-based substitution takes place). We observe D_j for products $j \in S$ from sales data. Notice that at a store that has full assortment (i.e., $S = N$), no substitution takes place, hence $D_j = d_j$ for all j. We can therefore estimate d_j for $j \in N$ from sales data of a similar store that carries a full assortment. We can conclude that the substitution rate is positive for this subcategory if $\Sigma_{j \in S} D_j > \Sigma_{j \in S} d_j$. Let $y(S) = \Sigma_{j \in S} D_j$. Given $\mathbf{d}$, substitution rate δ, and assortment S, we compute what each product in S would have sold at this store using Eq. (8.6), and the total subcategory sales denoted (S, δ). The error associated with a given δ is the difference between the observed and theoretical subcategory sales at a store [i.e., $y(S) - \widehat{y}(S, \delta)$]. We find the substitution rate δ that minimizes the total error across all available data from multiple stores and different time periods. The details of the procedure can be found in the paper.

As Campo et al. (2004) point out, there are significant similarities in consumer reactions to a permanent assortment reduction and to stockouts. Therefore, the substitution rate estimated for assortment based substitution can be also used for stockout-based substitution if that cannot be estimated. Another advantage of this methodology is that it enables us to estimate the demand rates of products in a store including those that have never been carried in that particular store.

The next step after the estimation of the substitution rate is the computation of the true demand rates. This involves two tasks. (a) deflating the demand rate of the variants already in the assortment S_h, and (b) estimating a positive demand rate for the variants that are not in S_h. Clearly, if $S_h = N$, no computation is necessary. Figure 8.1 presents an example of observed demand rates and the computed true demand rates for a subcategory with ten products.

5.3 Estimation of Non-parametric Choice Models

Farias et al. (2013) study the problem of modeling consumer choice, when the amount of data available is limited. They show that optimizing the assortment based on a mis-specified choice model can lead to highly suboptimal revenues. They consider a generic consumer choice model, where choices are modelled as

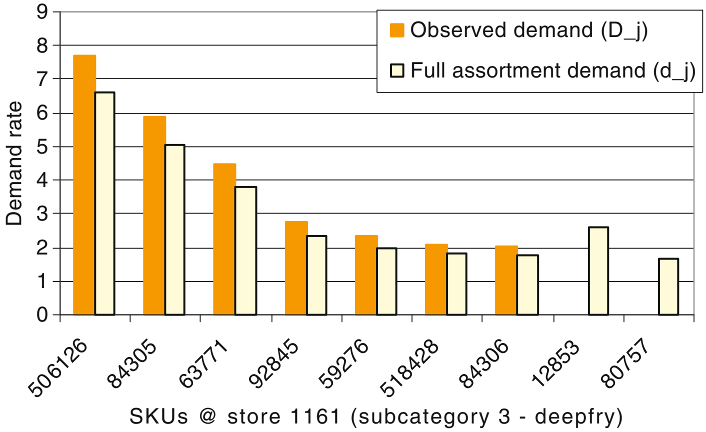


Fig. 8.1 Estimates of observed and original demand rates for a subcategory

distributions over preference lists. They develop a non-parametric approach to learn the right choice model, using limited data on customer purchase decisions. They apply their method on a real data set consisting of automobile sales transactions from a major US automaker, and show that it leads to a 20 % improvement in prediction accuracy over other state-of-the-art models, which results in a 10 % increase in revenues. They addresses the crucial issue of choice model identification, which is key to optimizing the assortment.

van Ryzin and Vulcano (2013) extend their previous work to estimate demand for a set of substitutable products using readily-available sales transactions and product availability data. They model demand as consisting of bernoulli arrivals followed by a general, non-parametric discrete choice model, that is compatible with an arbitrary random utility model. They apply the EM algorithm to jointly estimate the arrival rates and the probability distribution of customer choices. They use numerical experiments to demonstrate that their approach allows them to rapidly identify customer types and produce good estimates of demand.

6 Assortment Planning in Practice

The goal of this section is to describe assortment planning practice as illustrated by the processes used by a few retailers with whom we have interacted: Best Buy, Borders Books, Tanishq and Albert Heijn (Levy and Weitz 2004, Chapter 12), also provides a description of retail assortment planning.

6.1 Best Buy

Most retailers divide their products into various segments, usually called categories and sub categories. The assortment planning process begins by forecasting the sales of each segment for a future planning period ranging from a several month season to a fiscal year. Then scarce store shelf space and inventory purchase dollars are allocated to each segment based in part on the sales projections. Finally, given these resource allocations, the number of SKUs to be carried in each segment is chosen. As such, assortment planning in practice is essentially a strategic planning and capital budgeting process.

Best Buy offers a good example of this process. In their planning process, conventional still cameras and digital still cameras are two of the product segments. The starting point for a forecast of next year's sales is last year's sales adjusted for trend. Figure 8.2 shows sales of digital and traditional cameras through 2002. A logical forecast for 2003 would be less than 2002 sales for traditional cameras and more than 2002 sales for digital cameras.

The forecasts based on sales history are then adjusted based on information from trade shows, vendors, observations of competitor moves and reviews of new technology. The goal of assimilating these inputs is to identify changes in sales for a product category that might not be apparent from a straight forward extrapolation of sales history.

The next step is to set goals for each segment for sales, margin and market share based on the sales forecast, to allocate shelf space and inventory purchase dollars and then to determine how many SKUs to carry in each product segment. A critical input in deciding how many SKUs to carry is the importance to the customer of a

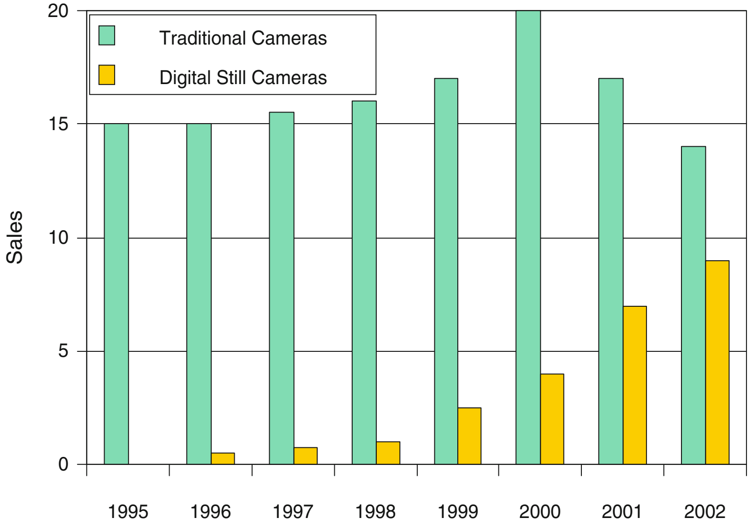
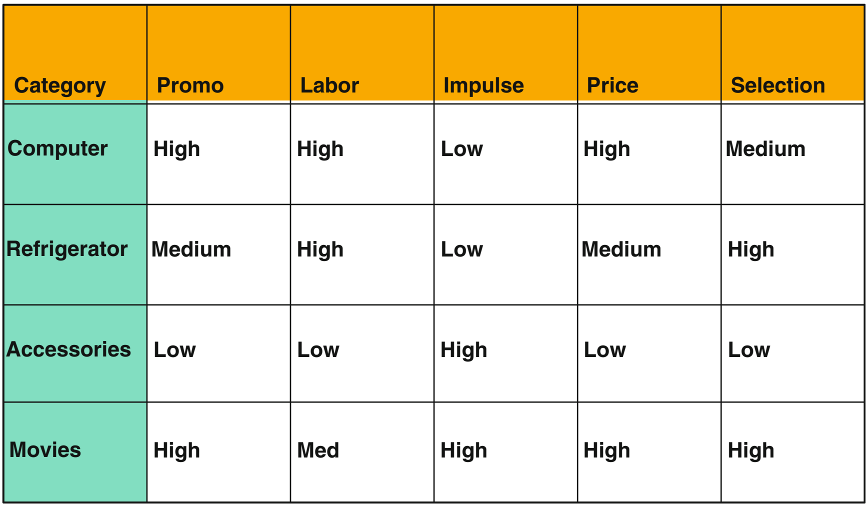


Fig. 8.2 Historical sales of traditional and digital cameras



Category	Promo	Labor	Impulse	Price	Selection
Computer	High	High	Low	High	Medium
Refrigerator	Medium	High	Low	Medium	High
Accessories	Low	Low	High	Low	Low
Movies	High	Med	High	High	High

Fig. 8.3 The impact of sales drivers for various types of products

broad selection in a particular category. Figure 8.3 was created by Best Buy to show the factors that influence sales and the importance of these factors for different types of products. For example, an accessory item such as a surge protector is often an impulse buy whose sales would be significantly increased by placing it on display near the check out register or in some other high traffic area. However, the customer is not particularly sensitive to price and doesn't require a broad selection. By contrast, placing a refrigerator next to the cash register to drive sales would be silly, because this isn't an impulse purchase for customers. However, they do value a broad selection and low prices. Another way of interpreting the data in this table is that Best Buy believes customers shopping for accessories are very willing to substitute if they don't find exactly what they are looking for, but refrigerator and movie customers are relatively unwilling to substitute.

This matrix is used to guide the number of SKUs to be carried in each product category. Other things being equal, a greater number of SKUs would be carried for those products where selection has a high impact on sales.

Once the number of SKUs to be carried in a product segment has been determined, it is left to the buyer for that segment to determine exactly which SKUs to carry. As an example, in flat panel TV's, Best Buy might carry 82 different SKUs. By contrast, the number of potential SKUs is much larger, comprising of eight diagonal widths (e.g. 19″, 25″, 32″, 35″, 40″, etc.), five screen types (plasma, LCD, projection, etc.), seven resolutions (analog, 480i, 720p, 1080i, etc.) and nine major vendors (Sony, Panasonic, Pioneer, etc.) for a total of $8 \times 5 \times 7 \times 9 = 2,520$ potential SKUs. It is left to the buyer through a largely manual process to determine which 82 out of these 2,520 SKUs will be carried by Best Buy. The buyer incorporates a number of factors into the choice of SKUs. For example, it is highly desirable to carry products from several vendors so that Best Buy can benefit from competition when negotiating with vendors on price.

The Best Buy example suggests that practice and academic research are complementary, in that practice ends with delegating to the buyer the decision of which products to carry from the universe, and this is precisely the problem that has been emphasized in the academic literature.

6.2 Borders

Two interrelated issues in assortment planning are the division of decision rights between corporate and stores and the degree to which the assortment varies by store. Figure 8.4 below depicts alternatives of these two factors.

By far the most common approach is for corporate headquarters to decide on a single common assortment that is carried by all stores of the chain, except that in smaller stores, the breadth of the assortment may be reduced by removing some of the least important SKUs. A relatively small number of retailers (Bed Bath & Beyond would be an example) allow their store managers considerable authority in deciding which SKUs to carry in their stores. Usually, a portion of the assortment is dictated by corporate, and the remainder is chosen by store management from a corporate approved list of options. Obviously a result of this approach is that the assortment is different in all stores, and is hopefully tuned to the tastes of that store's customers.

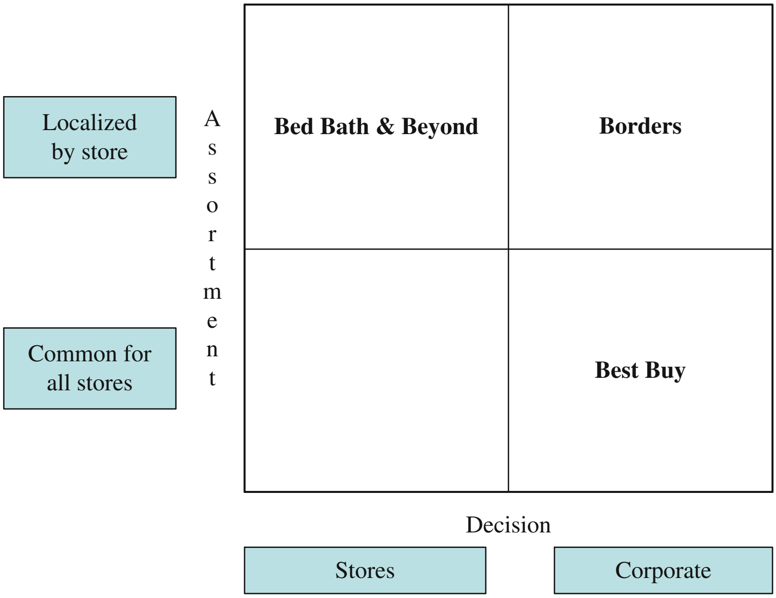


Fig. 8.4 Approaches to assortment planning

Borders Books is one of the few retailers that have developed a central approach to creating a unique assortment for each store. They segment their products into about 1,000 book categories and define the assortment at a store by the number of titles carried in each category. To choose these parameters they rely on a measure called Relative Sales per Title (RST) that equals the sales in a category over some history period divided by the number of titles carried in the category over the same period. If RST is high for a store-category in a recent period, then they increase the number of titles in that category, and conversely, reduce the titles in low RST categories. For example, a rule could be to divide the 1,000 categories in a store into the upper, middle and lower third of RST values and then increase number of titles carried in upper third by Δ and reduce lower third by Δ, where Δ and the frequency of adjusting the assortment are parameters of the process that determine how quickly and aggressively the assortment is adjusted based on history. Their overall process also takes seasonality into account, but that is outside the scope of this survey article.

6.3 Tanishq

Tanishq, a division of Titan Industries Ltd. (India's largest watch maker) is India's leading branded jewelry manufacturer and retailer in the country's $10 billion jewelry market. Tanishq jewelry is sold exclusively through a company controlled retail chain with over 60 boutiques spread over 39 cities. This network of boutiques is supplied and supported by a strong distribution network.

Assortment planning is a key activity at Tanishq involving significant challenges. First, jewelry is a complex product category with a very broad offering to choose from (more than 30,000 active SKUs) making assortment selection non-trivial. Second, given the small to medium size of most of the retail outlets, there were inventory limitations; as a consequence, getting the assortment decision right was critical. Significant differences in customer profile across its 60 boutiques and the frequent introduction of new products added further layers of complexity to the assortment planning process.

Traditionally, each store placed its own order, subject to guidelines on total inventory drawn up by the supply chain team at the corporate headquarters. This was done since the store associates were the ones closest to the customers and hence believed to have the best understanding of their preferences. This was true to a large extent, as the jewelry buying process in the Indian market was highly interactive, with store associates playing a significant role in guiding the customer through the product offerings based on their preferences (e.g. price range, design). Consequently, the store associates had a fairly accurate knowledge of customer choices, their willingness to substitute across product attributes, and reasons that led them to reject certain product variants.

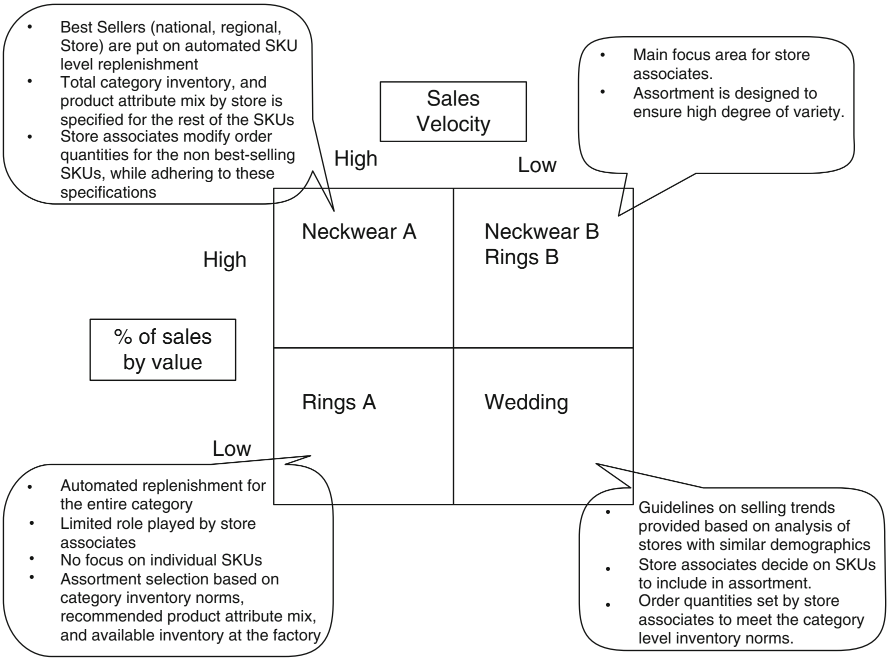
However, there were issues with this model. First, store associates were already burdened with monthly sales targets and hence had little time to do full justice to the ordering process. Second, their knowledge was limited only to product variants that the store had stocked in the past. Hence, they were missing out on potential product opportunities. This necessitated the need to modify the existing assortment planning process and address those shortcomings.

Tanishq accomplished this by moving from a store-centric model to a hybrid model involving both the store associates and a central supply chain team. The supply chain team at the corporate headquarters had the best access to sales and inventory data from all stores. They had detailed information about market trends and were in the best position to analyze historical data to detect selling patterns, and best selling variants at the state, regional, and national levels. This, combined with the local, store specific knowledge of the store associates, resulted in a more refined process for Tanishq.

The first step was the identification of product attributes relevant to the customers' choice process. This was done by the central supply chain team, based on inputs from the store associates. For example, the product category of rings was defined by the following attributes: theme, collection, design, gem type and size.

The next step was the determination of an appropriate assortment strategy for each product category. Again, this was carried out by the central supply chain team. They analyzed historical sales and inventory data in order to understand differences in sales mix across stores by attribute, to identify best sellers, and to develop an understanding of basic selling patterns.

The assortment strategy for each product category was developed based on a simple 2 × 2 matrix of percent contribution to sales vs. sales velocity (see Fig. 8.5).



- Best Sellers (national, regional, Store) are put on automated SKU level replenishment
- Total category inventory, and product attribute mix by store is specified for the rest of the SKUs
- Store associates modify order quantities for the non best-selling SKUs, while adhering to these specifications

- Main focus area for store associates.
- Assortment is designed to ensure high degree of variety.

- Automated replenishment for the entire category
- Limited role played by store associates
- No focus on individual SKUs
- Assortment selection based on category inventory norms, recommended product attribute mix, and available inventory at the factory

- Guidelines on selling trends provided based on analysis of stores with similar demographics
- Store associates decide on SKUs to include in assortment.
- Order quantities set by store associates to meet the category level inventory norms.

Fig. 8.5 Assortment strategy based on percent sales vs. sales velocity matrix

For example, in the case of a product category like Daily Neckwear, which has high percent sales contribution as well as high velocity, the high volume SKUs were put on replenishment, with inventory levels decided based on simple EOQ models. For the rest of the category, norms were drawn for overall inventory level and product attribute mix at each store (e.g. at Store A, overall inventory of Neckwear should be $ 2 million and the mix should be: Themes—50 % traditional, 30 % contemporary, and 20 % fashion; Gem—40 % large, 30 % medium, and 30 % small).

Based on the assortment strategy, the supply chain team developed a preliminary assortment plan for each store, with suggested products and inventory levels. With a bulk of the products put on SKU level replenishment, the work of store associates has been considerably reduced.

For the products not on SKU-level replenishment, the store associates were at liberty to modify the products selected and order quantities based on their knowledge of localized customer preferences. This was subject to the overarching inventory and product attribute mix guidelines drawn by the central team. This is done through a visual interface that provides the store associates a dynamic picture of how the modified order is stacking up against corporate guidelines.

Through the adoption of a hybrid model, Tanishq was thus able to customize its product offering to suit each store's clientele, while at the same time automating a bulk of the assortment planning process.

6.4 *Albert Heijn*

Albert Heijn, BV is a leading supermarket chain in the Netherlands with 1,187 stores and about $10 billion in sales.[2] In the grocery industry, supermarkets often carry more than 30,000 stock keeping units (SKUs). At the top level of the hierarchy, SKUs are divided into three groups: chilled products, dry goods, and groceries. Each group then is divided into merchandising categories, such as wines, bread spreads, butter and margarine. A subcategory is defined as a group of variants such that the difference between products within a subcategory is minimal, but the difference between subcategories is significant. For example, the subcategories in the butter and margarine category include deep-fry fat, regular butter, healthy butter, and margarines. We assume that substitution takes place within a subcategory but not across subcategories. The assortment planning models reviewed in this chapter focused on the selection and inventory/space allocation within a subcategory given a fixed shelf space and other constraints. Albert Heijn follows a hierarchical approach to assortment planning. First, store space is allocated to categories. Then product selection and facing allocation to products are

[2] Albert Heijn, BV is a subsidiary of Ahold Corporation, which owns many supermarket chains around the world with about 8,500 stores and $50 billion in sales.

carried out, subject to the shelf space constraint. In this subsection, we describe the details of this hierarchical approach.

Albert Heijn solves the following optimization problem to allocate shelf space between categories for each store.

$$\max\left\{\sum\nolimits_i P_i(x_i) : \sum\nolimits_i x_i \leq StoreShelfSpace;\ x_i \geq 0, \forall i.\right\}$$

$P_i(x_i)$ is the category gross profit when x_i meters of shelf space is allocated to category i. The function P_i is assumed to have a logarithmic form whose parameters are estimated using data from multiple stores $(x_i, P_i(x_i))$. The optimization is done by a Greedy Heuristic—allocating 1 m of shelf space at each step to the category with the highest incremental gross profit. Note that this shelf space allocation approach is similar to Corstjens and Doyle (1981), except that cross-space elasticities are not included in the formulation (i.e., category gross profit depends only on the category shelf space).

(Contrast this with the shelf space allocation approach at Borders Bookstores. Borders grouped 300,000 titles into 300 categories and allocated shelf space to categories on the premise that, "Except for best sellers, a customer is interested not in title but category". Category popularity is assessed by computing RST (Relative sales per title = Category sales/Number of titles). Shelf space is periodically reassigned from low RST to high RST. Following the principle of "Survival of the Fittest", categories "fight" for shelf space. Store managers are allowed to pick titles to be stocked within each category, thereby decentralizing a part of the decision process. Assuming that the number of titles is a proxy for category shelf space, RST is equivalent to $P_i(x_i)/x_i$. The Borders approach is similar to that of Albert Heijn except that rather than allocating the last meter of shelf space based on the marginal return, Borders allocates space based on average return from a category.

At Albert Heijn, it is the category manager's responsibility to choose the number of products and their shelf space allocation in each category, given a fixed shelf space. Category managers use several heuristics and their expertise about the category in order to make these decisions. Firstly, Albert Heijn wants to be known as the high variety, high quality supermarket in the Netherlands. One of the guidelines to achieve this strategical mandate is to carry 10 % more variety than the nearest competitor. The minimum number of SKUs in a subcategory, the minimum number of facings in a subcategory, the minimum and maximum number of facings for particular SKUs are also specified by category managers. If there is a need to reduce variety in a subcategory, the likely candidate is the subcategory with the highest substitution rate. To introduce new products periodically, m worse products are discarded and m new products are included in the assortment. Given the product selection, facings are allocated to products proportional to their demand rates.

Inventory management operates within the given facing allocations for a selection of products. For non-perishable items, the assigned facings are filled as much as possible at all times, even in the non-peak-load periods. That is achieved by ordering an integral number of case packs such that the inventory position is as close as possible to and less than the maximum inventory level that would fit in the allocated facings. For perishable items that have a shelf life of a few days or less (e.g., produce), the inventory control is done in a more dynamic way. Albert Heijn uses a real-time system that estimates the demand for each product in the assortment based on the sales in the last few hours, and places an order to maximize each product's expected revenues minus cost of disposed inventory.

6.5 Comparison of Academic and Industry Approaches to Assortment Planning

This section compares and contrasts the approaches taken by academia and industry to assortment planning. Industry has taken a more strategic and holistic approach, while academics use a more operational and detail oriented approach. In some respects these approaches are nicely complementary in that the aspects of assortment planning that have received least attention in practice have received the most attention in academia, and academic research has the potential to fill a void in retail practice.

For most retailers, the process of assortment planning starts at the strategic level. The breadth of product categories carried and the depth of products offered in each of them is a function of the retailer's position in the competitive landscape. For example, a retailer like Best Buy would carry a rarely demanded product such as a 10 mega-pixel camera, just to maintain consumer perception of Best Buy as offering the latest technologies. In other words, the assortment would carry products which are otherwise unprofitable, but are a strategic necessity. While academic research does acknowledge such phenomenon (Cachon et al. 2005), there is little research that focuses on incorporating these strategic considerations while optimizing the assortment.

The other strategic aspect that retailers are concerned with is the role of a product category in their mix. Going back to the Best Buy example, it might be the case that Best Buy offers a very extensive assortment of HDTV's, more than what might be the optimal number when looked upon in isolation, for they are the main traffic drivers for the store. In other words, customers prefer to shop at Best Buy as they see extensive variety on offer in key categories, and as a result end up buying at Best Buy. There is little academic research (except Cachon and Kök 2007) that models this aspect of an assortment. On the other hand, the pricing version of this phenomenon (loss leaders and advertising features to drive traffic into the store and the razor-blade model) is extensively studied in the marketing literature.

One common theme across all the industry examples is that retailers recognize the fact that not all categories should be treated the same. The major drivers of sales in each category are different. While product variety may be the most important factor in a consumers store choice and purchasing decisions for one category, promotions, in-store service experience, and impulse buying (aisle displays) may be more critical for another category. For example, Dhar et al. (2001) find that increasing the breadth and depth of the assortment does not have a positive effect on the performance of high penetration, high frequency categories like coffee and cereals.

Most retailers consider product selection as one among several levers (like promotions, pricing, etc.) that influence sales. Hence, they find it critical to integrate assortment planning decisions with the other influencing parameters. For example, if an apparel retailer is advertising a certain line of clothing heavily, then the variety that needs to be offered is higher than what might have been required without the attention due to advertising. Hence, retailers make assortment decisions in conjunction with other key factors that influence sales.

Retailers are well aware of the dynamic nature of the problem. At many retailers, the initial assortment developed by the buyers is tested across a sample of stores to get an early read, prior to the actual selling season. The test results are used to understand trends on winners and losers and gaps in the portfolio so as to redesign the assortment. As there are several other factors such as promotions, pricing, display, etc. which affect sales on an ongoing basis, the assortment is reviewed from time to time and appropriate changes are made. Academic papers, with the exception of Caro and Gallien (2007), consider static assortments. Even in mature categories, the frequent introduction of new products make it a necessity to revise the assortments. In practice, categories in different stages of their life cycles or categories with seasonal products require different assortment planning approaches. Growth potential is another strategic consideration that influences a retailer's assortment. For example, a dying product category like VCRs might not have the variety that a growing category like DVDs would.

The Tanishq example illustrates how assortment planning and replenishment can be attribute-focused rather than product-focused. For non-best sellers, Tanishq chooses a certain theme and gem size distribution as the defining properties of the target assortment. This approach is sensible, especially for categories in which attributes of the products are critical in driving traffic and influencing consumers' choice behavior. The attribute-focused approach is common in apparel retailing as well. Levy and Weitz (2004) describe the assortment plan for a jeans category where the size distribution, colors and styles are the main attributes that define the assortment. The total inventory budget is then allocated to products given the required distribution of the assortment over these attributes. Academic assortment planning models are mostly product-focused.

Customization of the assortment at the store level has gotten scant attention from retailers and no attention from academics. The Tanishq example illustrates a hybrid approach, where either the assortment or the guidelines for the assortment of the categories are planned at the corporate level, and for some categories store associates tinker with the assortment given the guidelines. Albert Heijn also follows the hybrid approach in that the store assortments are chosen from a chain-wide assortment. Borders Books is the best example we know of a retailer that aggressively customizes assortments at the store level.

Retailers take supply chain considerations into account in assortment planning. For example, Best Buy considers vendor relations, vendor performance and the number of products in other categories from a vendor while developing the assortment plans. However, there is very limited discussion of assortment planning from a supply chain view in the academic literature.

We performed a search on Google for "retail assortment planning" and found more than 700 references. Most of these references are to the product description of software providers and consulting firms, indicating a strong industry interest in the topic. Some academic papers come up in the search as well. One interesting observation that complements the discussion above is that there is a huge disconnect between the two groups: the language or the terminology of each group is substantially different and neither group acknowledges the existence of the other.

7 Directions for Future Research

There has been strong interest in assortment planning research since the first edition of this book chapter in 2008. Four research avenues emerge as important future research directions based on our discussion in this chapter.

First, more empirical work is needed in understanding the impact of assortment variables on consumers' store choice and purchasing behavior. Second, most of the existing theoretical models have not been implemented as part of industry applications (or their theoretical predictions have not been empirically tested). The field would benefit from such applications and empirical tests, as a validation of the assumptions in the increasingly complicated assortment planning models being formulated in the academic literature. Third, it seems that there are significant opportunities in generalizing the existing theoretical work to handle more complex problems faced by the retailers. One example would be to allow customization of the assortment by store. Fourth, incorporating the empirical findings on consumer behavior and perception of variety in assortment optimization models seems a worthy area of research. Below we describe some possible research topics from these four avenues in no particular order.

Demand arrival is assumed to be exogenous in most academic models. Understanding the drivers of store traffic through market share or store choice models, and incorporating those in assortment planning is a possible research direction. Lower prices, for example, would increase store traffic, but on the other hand, lower margins would lead to narrower assortments. Retailers recognize these interactions but make these decisions sequentially and in rudimentary ways. The joint pricing and assortment planning problem has not been studied in depth. Aydin and Ryan (2000) study optimal pricing under MNL model but do not consider operational costs. Cachon et al. (2008) are interested in the impact of competitive intensity on the variety level and prices.

Academic models take a static view of the assortment planning problem, whereas in practice, assortment decisions in a category can be made several times throughout the season. The problems that industry faces include not only multi-period problems, but also managing the assortment for multiple generations of products, as in the digital versus traditional camera example. The dynamic assortment problem provides a rich set of research questions.

A significant number of papers have started studying dynamic assortment planning. Demand learning through tests in sample stores or online environments remain a topic worthy of investigation. Online retail environments and omni-channel retailing bring up many novel applications of dynamic assortment planning and open research questions.

Assortment planning models assume that there is a well defined set of candidate products, for which the consumer choice behavior is known perfectly. It may be interesting to take an attribute view of this problem, where consumers are interested in particular attributes rather than products. Mostly, a category is assumed to be composed of homogenous products that are potential substitutes from a consumer's perspective. Assortment planning for vertically differentiated products (i.e., varying quality) or more general choice models (e.g., subgroups of products that are more likely than others to be substitutes) can be studied to generalize the existing results on properties of optimal assortments. There is a significant body of literature in marketing on consumers' perception of variety as mentioned in Sect. 2.4. Incorporating some of those concepts in assortment planning may increase the applicability of the theoretical models.

Consumers are usually assumed to be a homogenous group. However, marketing literature places particular emphasis on understanding consumer segments. Estimation papers attempt to identify the latent consumer segments, and products are carefully positioned to achieve price discrimination between consumer segments in the product line design literature. Similarly in retail assortment planning, the consideration of multiple consumer segments may lead to optimal assortments that are composed of clusters of products that target these different segments. Recent work on mixed logit models and assortment customization provide a starting point in this direction.

Consumer purchase decisions across product categories may not always be independent. For example, a consumer's decision to buy a red colored sheet might depend on his being able to find a matching pillow. Explicitly incorporating this basket effect of consumer behavior while optimizing the assortment is an interesting research avenue. Agrawal and Smith (2003) and Cachon and Kök (2007) are first examples of this.

Estimating model parameters such as substitution probabilities, is another area that needs further research. There is an extensive body of literature in marketing (conjoint analysis) and econometrics that deal with parameter estimation for a wide variety of consumer choice models. However, there is little application of these in the assortment planning literature. For academic research to impact the industry, it is critical to invest research time in this area and to come up with innovative techniques to estimate the parameters which form the backbone of the several optimization models.

It is usually assumed that each individual buys a single unit of a single product in a category. This may not be true, even among substitutable products. For example, one shopper may buy multiple units of multiple flavors of yogurt in the same purchase occasion. This behavior violates the assumptions of standard choice models like the MNL, and it might be interesting to develop alternate models and study the properties of the resulting assortment. It would also be worthwhile to study the structure of the optimal assortment for product categories in situations when consumers are variety-seeking, causing the inventory-variety trade-off to take a different form.

Clearly, it is necessary to develop methods to understand the role of categories and to measure the intangible factors (such as the strategic importance of a category, the impact of assortment breadth or inventory levels on attractiveness of a store). The relation of assortment and inventory decisions with other levers such as pricing, promotions, and advertising has not been studied empirically. Joint optimization of some of these variables may lead to interesting results. It may be possible to draw from the literature on economics of product differentiation and the marketing/ operations literature on product line design, both of which have extensively studied these variables and their impact on industry structures or product variety.

Assortment planning in multi-store, multi-tier supply chains is a completely open research area. Singh et al. (2005) and Aydin and Hausman (2003) are the only cases in the literature that incorporate supply chain considerations into assortment planning. The pros and cons of the hierarchical approach, the benefits of localization, and the execution problems associated with them have not been studied empirically or analytically. Balancing the benefits of customizing assortments by store with the increased cost of complexity is increasingly seen by retailers as a significant source of competitive advantage. An extremely interesting research question here is how to strike the balance, find the sweet spot between a "one size fits all" and "each store is its own" philosophies.

Incentive conflicts between the levels of the hierarchy may be a hurdle in deployment of the corporate assortment plans to the store level. Corporate level plans that are built based on strategic considerations may be imperfectly executed because the store managers' incentives are based on more short term objectives. The conflict of incentives between store managers, buyers, and vendors in a decentralized supply chain is yet another potential research area. For example, it is not clear how a category level assortment plan and the vendor-managed inventory agreements should be reconciled.

In conclusion, it seems to us that academics could make a tremendous contribution to retailing in the area of assortment planning. Retailers have developed practices that enable them to incorporate the complexities of the world in which they live, but they realize their approaches are too much based on art and judgment and that they could benefit from more rigorous use of the huge quantities of data available to them. If academics would be willing to work with individual retailers to understand their true complexity, they could make an enormous contribution in adding rigor and science to the retailer's planning process, much as academics have done in other areas like finance, marketing and strategy.

References

AC Nielsen. (1998). *Eighth annual survey of trade promotion practices*. Chicago, IL: ACNielsen.

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

Agrawal, N., & Smith, S. A. (2003). Optimal retail assortments for substitutable items purchased in sets. *Naval Research Logistics, 50*(7), 793–822.

Alptekinoğlu, A. (2004). Mass customization vs. mass production: Variety and price competition. *Manufacturing & Service Operations Management, 6*(1), 98–103.

Alptekinoğlu, A., & Grasas, A. (2014). When to carry eccentric products? Optimal retail assortment under consumer returns. *Production and Operations Management, 23.5*, 877–892.

Alptekinoğlu, A., Honhon, D., & Ulu, C. (2012). Positioning and pricing of horizontally differentiated products. Available at SSRN 2166570.

Alptekinoğlu, A., & Semple, J. (2013). *The exponomial choice model*. Working Paper, Pennsylvania State University.

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

Anupindi, R., Dada, M., & Gupta, S. (1998). Estimation of consumer demand with stockout based substitution: An application to vending machine products. *Marketing Science, 17*, 406–423.

Avsar, Z. M., & Baykal-Gursoy, M. (2002). Inventory control under substitutable demand: A stochastic game application. *Naval Research Logistics, 49*, 359–375

Aydin, G., & Hausman, W. H. (2003). *Supply chain coordination and assortment planning*. Working Paper, University of Michigan.

Aydin, G., & Ryan, J. K. (2000). Product line selection and pricing under the multinomial logit choice model. In *Proceedings of the 2000 MSOM Conference*.

Bassok, Y., Anupindi, R., & Akella, R. (1999). Single-period multiproduct inventory models with substitution. *Operations Research, 47*, 632–642.

Basuroy, S., & Nguyen, D. (1998). Multinomial logit market share models: Equilibrium characteristics and strategic implications. *Management Science, 44*(10), 1396–1408.

Baumol, W. J., & Ide, E. A. (1956). Variety in retailing. *Management Science, 3*, 93–101.

Bell, D. R., Ho, T.-H., & Tang, C. S. (1998). Determining where to shop: Fixed and variable costs of shopping. *Journal of Marketing Research, 35*, 352–369.

Bell, D. R., & Lattin, J. M. (1998). Shopping behavior and consumer preference for store price format: Why large basket shoppers prefer EDLP. *Marketing Science, 17*, 66–88.

Ben-Akiva, M., & Lerman, S. R. (1985). *Discrete choice analysis: Theory and application to travel demand*. Cambridge, MA: The MIT Press.

Bernstein, F., Gürhan Kök, A., & Xie, L. (2011). *Dynamic assortment customization with limited inventories*. Working Paper, Duke University.

Besbes, O., & Saure, D. (2011). *Product assortment and price competition with informed consumers*. Working Paper, Columbia University.

Boatwright, P., & Nunes, J. C. (2001). Reducing assortment: An attribute-based approach. *Journal of Marketing, 65*(3), 50–63.

Borin, N., & Farris, P. (1995). A sensitivity analysis of retailer shelf management models. *Journal of Retailing, 71*, 153–171.

Broniarczyk, S. M., Hoyer, W. D., & McAlister, L. (1998). Consumers' perception of the assortment offered in a grocery category: The impact of item reduction. *Journal of Marketing Research, 35*, 166–176.

Bucklin, R.E., & Gupta, S. (1992). Brand choice, purchase incidence, and segmentation: An integrated modeling approach. *Journal of Marketing Research, 29*, 201–215.

Bultez, A., & Naert, P. (1988). SHARP: Shelf allocation for retailers profit. *Marketing Science, 7*, 211–231.

Cachon, G. P., & Kök, A. G. (2007). Category management and coordination of categories in retail assortment planning in the presence of basket shoppers. *Management Science, 53*(6), 934–951.

Cachon, G. P., Terwiesch, C., & Xu, Y. (2005). Retail assortment planning in the presence of consumer search. *Manufacturing & Service Operations Management, 7*(4), 330–346.

Cachon, G. P., Terwiesch, C., & Xu, Y. (2008). On the effects of consumer search and firm entry in a multiproduct competitive market. *Marketing Science, 27.3*, 461–473

Campo, K., Gijsbrechts, E., & Nisol, P. (2004). Dynamics in consumer response to product unavailability: Do stock-out reactions signal response to permanent assortment reductions? *Journal of Business Research, 57*, 834–843.

Caro, F., & Gallien, J. (2007). Dynamic assortment with demand learning for seasonal consumer goods. *Management Science, 53.2*, 276–292.

Chen, F., Eliashberg, J., & Zipkin, P. (1998). Customer preferences, supply-chain costs, and product line design. In T.-H. Ho & C. S. Tang (Eds.), *Product variety management: Research advances*. Norwell: Kluwer Academic Publishers.

Chiang, J. (1991). A simultaneous approach to the whether, what and how much to buy questions. *Marketing Science, 10*, 297–315.

Chintagunta, P. K. (1993). Investigating purchase incidence, brand choice and purchase quantity decisions of households. *Marketing Science, 12*, 184–208.

Chong, J. K., Ho, T. H., & Tang, C. S. (2001). A modeling framework for category assortment planning. *Manufacturing & Service Operations Management*, **3**(3), 191–210.

Cooper, L. G., & Nakanishi, M. (1988). *Market-share analysis: Evaluating competitive marketing effectiveness*. Boston: Kluwer Academic Publishers.

Corstjens, M., & Doyle, P. (1981). A model for optimizing retail space allocations. *Management Science, 27*, 822–833.

Davis, J. M., Guillermo, G., & Topaloglu, H. (2014). Assortment optimization under variants of the nested logit model. *Operations Research, 62*(2), 250–273.

de Groote, X. (1994). Flexibility and marketing/manufacturing coordination. *International Journal of Production Economics, 36*, 153–167.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B, 39*, 1–38.

Desai, P., Radhakrishnan, S., & Srinivasan, K. (2001). Product differentiation and commonality in design: Balancing revenue and cost drivers. *Management Science, 47*, 37–51.

DeHoratius, N., & Raman, A. (2008). Inventory record inaccuracy: An empirical analysis. *Management Science, 54.4*, 627–641.

Dhar, S. K., Hoch, S. J., & Kumar, N. (2001). Effective category management depends on the role of the category. *Journal of Retailing, 77*(2), 165–184.

Dobson, G., & Kalish, S. (1993). Heuristics for pricing and positioning a product line. *Management Science, 39*, 160–175.

Downs, B., Metters, R., & Semple, J. (2002). Managing inventory with multiple products, lags in delivery, resource constraints, and lost sales: A mathematical programming approach. *Management Science, 47*, 464–479.

Dreze, X., Hoch, S. J., & Purk, M. E. (1994). Shelf management and space elasticity. *Journal of Retailing, 70*, 301–326.

Eliashberg, J., & Steinberg, R. (1993). Marketing-production joint decision-making. In J. Eliashberg & G. L. Lilien (Eds.), *Handbooks in OR & MS* (Vol. 5). Amsterdam: Elsevier

Emmelhainz, L., Emmelhainz, M., & Stock, J. (1991). Logistics implications of retail stockouts. *Journal of Business Logistics, 12*(2), 129–141.

Fader, P. S., & Hardie, B. G. S. (1996). Modeling consumer choice among SKUs. *Journal of Marketing Research, 33*, 442–452.

Farias, V. F., Jagabathula, S., & Shah, D. (2013). A nonparametric approach to modeling choice with limited data. *Management Science, 59.2*, 305–322.

Fisher, M. L., & Raman, A. (1996). Reducing the cost of demand uncertainty through accurate response to early sales. *Operations Research, 44*, 87–99.

Fisher, M., & Vaidyanathan, R. (2014). A demand estimation procedure for retail assortment optimization with results from implementations. *Management Science 60*(10), 2401–2415.

Gaur, V., & Honhon, D. (2006). Assortment planning and inventory decisions under a locational choice model. *Management Science, 52*(10), 1528–1543.

Greene, W. H. (1997). *Econometric analysis*. Englewood Cliffs, NJ: Prentice Hall.

Gruca, T. S., & Sudharshan, D. (1991). Equilibrium characteristics of multinomial logit market share models. *Journal of Marketing Research, 28*(11), 480–482.

Gruen, T. W., Corsten, D. S., & Bharadwaj, S. (2002). Retail out-of-stocks: A worldwide examination of extent, causes and consumer responses. Grocery Manufacturers of America.

Guadagni, P. M., & Little, J. D. C. (1983). A logit model of brand choice calibrated on scanner data. *Marketing Science, 2*, 203–238.

Hadley, G., & Whitin, T. M. (1963). *Analysis of inventory systems*. Englewood Cliffs, NJ: Prentice Hall.

Hoch, S. J., Bradlow, E. T., & Wansink, B. (1999). The variety of an assortment. *Marketing Science, 18*(4), 527–546.

Honhon, D., Gaur, V., & Seshadri, S. (2010). Assortment planning and inventory decisions under stockout-based substitution. *Operations Research, 58.5*, 1364–1379.

Honhon, D., Jonnalagedda, S., & Pan, X. A. (2012) Optimal algorithms for assortment selection under ranking-based consumer choice models. *Manufacturing & Service Operations Management, 14.2*, 279–289.

Hopp, W. J., & Xu, X. (2008). A static approximation for dynamic demand substitution with applications in a competitive market. *Operations Research, 56.3*, 630–645.

Hotelling, H. (1929). Stability in competition. *Economic Journal, 39*, 41–57

Huffman, C., & Kahn, B. E. (1998). Variety for sale: Mass customization or mass confusion? *Journal of Retailing, 74*, 491–513.

Irion, J., Al-Khayyal, F., & Lu, J. (2012). A piecewise linearization framework for retail shelf space management models. *European Journal of Operational Research, 222*(1), 122–136.

Jain, A., Rudi, N., & Wang, T. (2014). Demand estimation and ordering under censoring: Stock–out timing is (almost) all you need. *Operations Research, 63*(1), 134–150.

Kahn, B. E. (1995). Consumer variety-seeking in goods and services: An integrative review. *Journal of Retailing and Consumer Services, 2*, 139–48.

Kohli, R., & Sukumar, R. (1990). Heuristics for product line design. *Management Science, 36*(3), 1464–1478.

Kök, A. G. (2003). *Management of product variety in retail operations*. Ph.D. Dissertation, The Wharton School, University of Pennsylvania.

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

Kök, A. G., & Xu, Y. (2011). Optimal and competitive assortments with endogenous pricing under hierarchical consumer choice models. *Management Science, 57.9*, 1546–1563.

Kök, A., & Martínez-de-Albéniz, V. (2013). *A Competitive Model for Quick–Response Product Decisions*. Working Paper, Duke University.

Kurt Salmon Associates. (1993). *Efficient consumer response: Enhancing consumer value in the grocery industry*. Food Marketing Institute Report # 9–526, Food Marketing Institute.

Kurtulus, M. (2005). *Supply chain collaboration practices in consumer goods industry*. Ph.D. Dissertation, INSEAD.

Kurtulus, M., & Toktay, B. (2007). *Category captainship: Outsourcing retail category management*. Working Paper, Vanderbilt University.

Lancaster, K. (1966). A new approach to consumer theory. *Journal of Political Economy, 74*, 132–57.

Lancaster, K. (1975). Socially optimal product differentiation. *American Economic Review, 65*, 567–585.

Lancaster, K. (1990). The economics of product variety: A survey. *Marketing Science, 9*, 189–210.

Levy, M., & Weitz, B. A. (2004). *Retailing management* (pp. 398–400). New York: McGraw-Hill/Irwin.

Li, Z. (2007). A single-period assortment optimization model. *Production and Operations Management, 16.3*, 369–380.

Lippman S. A., & McCardle, K. F. (1997). The competitive newsboy. *Operations Research, 45*, 54–65.

Maddah, B., & Bish, E. K. (2004). Joint pricing, assortment, and inventory decisions for a retailer's product line. 2007. *Naval Research Logistics, 54*(3), 315–330.

Mahajan, S., & van Ryzin, G. J. (1999). Retail inventories and consumer choice. Chapter 17. In S. Tayur, et al. (Eds.), *Quantitative methods in supply chain management*. Amsterdam: Kluwer.

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

Mahajan, S., & van Ryzin, G. (2001b). Inventory competition under dynamic consumer choice. *Operations Research, 49*(5), 646–657.

Manchanda, P., Ansari, A., & Gupta, S. (1999). The "shopping basket": A model for multicategory purchase incidence decisions. *Marketing Science, 18*(2), 95–114.

Martínez-de-Albéniz, V., & Roels, G. (2011). Competing for shelf space. *Production and Operations Management 20*(1), 32–46.

McBride, R. D., & Zufryden, F. S. (1988). An integer programming approach to the optimal product line selection problem. *Marketing Science, 7*(2), 126–140.

McFadden, D. (1974). Conditional logit analysis of qualitative choice behavior. In P. Zarembka (Ed.), *Frontiers in econometrics*. New York: Academic.

McGillivray, A. R., & Silver, E. A. (1978). Some concepts for inventory control under substitutable demand. *INFOR, 16*, 47–63.

Miller, C. M., Smith, S. A., McIntyre, S. H., & Achabal, D. D. (2010). Optimizing retail assortments for infrequently purchased products. *Journal of Retailing, 86*(2), 159–171

Miranda Bront, J., Mendez-Diaz, I., & Vulcano, G. (2009). A column generation algorithm for choice-based network revenue management. *Operations Research, 57*(3), 769–784.

Moorthy, S. (1984). Market segmentation, self-selection, and product line design. *Marketing Science, 3*, 288–307.

Musalem, A., et al. (2010). Structural estimation of the effect of out-of-stocks. *Management Science, 56.7*, 1180–1197.

Mussa, M., & Rosen, S. (1978). Monopoly and product quality. *Journal of Economic Theory, 18*, 301–317.

Netessine, S., & Rudi, N. (2003). Centralized and competitive inventory models with demand substitution. *Operations Research, 51*, 329–335.

Netessine, S., & Taylor, T. A. (2007). Product line design and production technology. *Marketing Science, 26*(1), 101–117.

Noonan, P. S. (1995). *When consumers choose: A multi-product, multi-location newsboy model with substitution*. Working Paper, Emory University.

Pan, X. A., & Honhon, D. (2012). Assortment planning for vertically differentiated products. *Production and Operations Management, 21.2*, 253–275.

Parlar, M. (1985). Optimal ordering policies for a perishable and substitutable product: A Markov decision model. *Infor, 23*, 182–195.

Parlar, M., & Goyal, S. K. (1984). Optimal ordering policies for two substitutable products with stochastic demand. *Opsearch, 21*(1), 1–15.

Progressive Grocer. (1968a, October). The out of stock study: Part I. S1–S16.

Progressive Grocer. (1968b, November). The out of stock study: Part II. S17–S32.

Quelch, J. A., & Kenny, D. (1994). Extend profits, not product lines. *Harvard Business Review, 72*, 153–160.

Rajaram, K. (2001). Assortment planning in fashion retailing: Methodology, application and analysis. *European Journal of Operational Research, 129*, 186–208.

Rajaram, K., & Tang, C. S. (2001). The impact of product substitution on retail merchandising. *European Journal of Operational Research, 135*, 582–601.

Raman, A., McClellan, A. d., & Fisher, M. L. (2001). Supply chain management at world Co. Ltd. Harvard Business School Case # 601072.

Rusmevichientong, P., & Topaloglu, H. (2012). Robust assortment optimization in revenue management under the multinomial logit choice model. *Operations Research, 60.4*, 865–882.

Rusmevichientong, P., Shen, Z.-J. M., & Shmoys, D. B. (2010). Dynamic assortment optimization with a multinomial logit choice model and capacity constraint. *Operations Research, 58.6*, 1666–1680.

Russell, G. J., Bell, D. R., et al. (1997). Perspectives on multiple category choice. *Marketing Letters, 8*(3), 297–305.

Saure, D., & Zeevi, A. (2013). Optimal dynamic assortment planning with demand learning. *Manufacturing & Service Operations Management, 15*(3), 387–404.

Schary, P., & Christopher, M. (1979). The anatomy of a stockout. *Journal of Retailing, 55*(2), 59–70.

Simonson, I. (1999). The effect of product assortment on buyer preferences. *Journal of Retailing, 75*, 347–370.

Singh, P., Groenevelt, H., & Rudi, N. (2005). *Product variety and supply chain structures*. Working Paper, University of Rochester.

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

Song, J.-S. (1998). On the order fill rate in multi-item, base-stock systems. *Operations Research, 46*, 831–845.

Song, J.-S., & Zipkin, P. (2003). Supply chain operations: Assemble-to-order systems. In S. Graves & T. De Kok (Eds.), *Handbooks in operations research and management science. Supply chain management* (Vol. XXX). North-Holland: Amsterdam.

Talluri, K., & van Ryzin, G. (2004). Revenue management under a general discrete choice model of consumer behavior. *Management Science, 50*, 15–33.

Ulu, C., Honhon, D., & Alptekinoğlu, A. (2012). Learning consumer tastes through dynamic assortments. *Operations Research, 60.4*, 833–849.

Urban, T. L. (1998). An inventory-theoretic approach to product assortment and shelf space allocation. *Journal of Retailing, 74*, 15–35.

Vaidyanathan, R., & Fisher, M. (2012). *Assortment planning*. Working Paper, The Wharton School, University of Pennsylvania.

van Herpen, E., & Pieters, R. (2002). The variety of an assortment: An extension to the attribute-based approach. *Marketing Science, 21*(3), 331–341.

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

van Ryzin, G., & Vulcano, G. (2013). *An expectation-maximization algorithm to estimate a general class of non-parametric choice-models*. Working Paper.

Vulcano, G., Van Ryzin, G., & Ratliff, R. (2012). Estimating primary demand for substitutable products from sales transaction data. *Operations Research, 60.2*, 313–334.

Walter, C., & Grabner, J. (1975). Stockout models: Empirical tests in a retail situation. *Journal of Marketing, 39*, 56–68.

Wu, C. F. J. (1983). On the convergence properties of the EM algorithm. *Annals of Statistics, 11*, 95–103.

Zinn, W., & Liu, P. (2001). Consumer response to retail stockouts. *Journal of Business Logistics, 22*(1), 49–71.