OPERATIONS RESEARCH
Vol. 55, No. 6, November–December 2007, pp. 1001–1021
ISSN 0030-364X | EISSN 1526-5463 | 07 | 5506 | 1001

informs

DOI 10.1287/opre.1070.0409

Demand Estimation and Assortment Optimization Under Substitution: Methodology and Application

A. Gürhan Kök
Fuqua School of Business, Duke University, Durham, North Carolina 27708, gurhan.kok@duke.edu

Marshall L. Fisher
The Wharton School, University of Pennsylvania, Philadelphia, Pennsylvania 19104, fisher@wharton.upenn.edu

Assortment planning at a retailer entails both selecting the set of products to be carried and setting inventory levels for each product. We study an assortment planning model in which consumers might accept substitutes when their favorite product is unavailable. We develop an algorithmic process to help retailers compute the best assortment for each store. First, we present a procedure for estimating the parameters of substitution behavior and demand for products in each store, including the products that have not been previously carried in that store. Second, we propose an iterative optimization heuristic for solving the assortment planning problem. In a computational study, we find that its solutions, on average, are within 0.5% of the optimal solution. Third, we establish new structural properties (based on the heuristic solution) that relate the products included in the assortment and their inventory levels to product characteristics such as gross margin, case-pack sizes, and demand variability. We applied our method at Albert Heijn, a supermarket chain in The Netherlands. Comparing the recommendations of our system with the existing assortments suggests a more than 50% increase in profits.

Subject classifications: inventory: multi-item, stochastic, applications, heuristics; marketing: retailing, estimation.
Area of review: Manufacturing, Service, and Supply Chain Operations.
History: Received October 2004; revisions received January 2006, August 2006; accepted September 2006. Published online in *Articles in Advance* October 15, 2007.

1. Introduction

Retail assortment planning is defined as specifying the set of products carried at each store and setting their inventory levels so as to maximize a profit function subject to fixed shelf space and possibly other constraints, which vary by context. Consumers sometimes cannot find their favorite product in a store and settle 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.

In this paper, we describe a novel methodology for estimating the input data-demand and substitution parameters for the assortment planning problem and propose an optimization algorithm. We applied our method at Albert Heijn, BV, a leading supermarket chain in The Netherlands with 1,187 stores and about $10 billion in sales. (Albert Heijn is a subsidiary of Ahold Corporation, which owns many supermarket chains around the world with about 8,500 stores and $50 billion in sales.) Although our methodology is applicable to more general cases, we describe the problem and our methodology in the context of Albert Heijn's operations for expositional simplicity.

The number of products carried in a retail store can be very large. In the grocery industry, supermarkets often carry more than 30,000 stock keeping units (SKUs). At the highest level of the hierarchy, SKUs are divided into three groups: chilled products, dry goods, and groceries. Each group is then divided into merchandising categories, such as wines, bread spreads, and butter & margarine. Within categories, subcategories are defined so that the difference between products within a subcategory is minimal, but the difference between subcategories is significant. For example, the subcategories of the butter & 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 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 a fixed lead time. There is no back room; therefore, orders are delivered directly 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. The 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 f_j allocated to product j determines its maximum level of inventory, $c_j f_j$, where c_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 from product to product. Perishable products are disposed at the end of their shelf life. The loss associated with disposing a unit is approximately

equal to the selling price. The objective is to maximize expected gross profit, where gross profit is defined as per-unit margin × sales − selling price × disposed inventory.

Because subcategories are defined so that there is minimal interaction between the products in different subcategories, we start with analyzing each subcategory separately. The decision process for a subcategory involves allocating a discrete number of facings to each product to maximize total expected gross profits subject to a shelf-space constraint:

$$\begin{aligned} \max_{f_j,\, j\in N} \quad & Z = \sum_j G_j(f_j, D_j(\mathbf{f}, \mathbf{d})) \\ \text{s.t.} \quad & \sum_j f_j w_j \leqslant \textit{shelf space}, \qquad \text{(AP)} \\ & f_j \in \{0, 1, 2, \ldots\} \quad \text{for all } j, \end{aligned}$$

where $N = \{1, \ldots, J\}$ denotes the set of potential variants in the subcategory, 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 to product j, depends on the facing allocation and the demand rates of all products in the subcategory, i.e., $\mathbf{f} = (f_1, f_2, \ldots, f_J)$ and $\mathbf{d} = (d_1, d_2, \ldots, d_J)$, where d_j is the original demand rate to product j (i.e., the number of customers who would select j as their first choice if presented with all J products). We assume that the original demand for a product is not affected by the number of facings assigned to it. The store's assortment is denoted S and is determined by the facing allocation, i.e., $S = \{j \in N\colon f_j > 0\}$.

The unknown inputs to the assortment optimization (AP) are the original demand rates d_j and the function D_j. There is plenty of empirical evidence suggesting that consumers might be willing to accept a substitute if their favorite product is unavailable (e.g., Fitzsimons 2000, Gruen et al. 2002, Campo et al. 2004). Hence, under substitution, the observed demand for each product would be different from its original demand. In other words, we do not observe d_j from store sales data. Rather, we observe the value of the D_j function given the current assortment $\mathbf{f}$. In §3, we describe our model of substitution. In §4, we present two methodologies for estimating demand for products and the parameters of the substitution model. The first method estimates demand and substitution rates using sales data from multiple stores when the service levels are high enough so that stockouts are negligible and only assortment-based substitution (from products that are not included in the assortment to the products that are in the assortment) takes place. The second method generalizes our approach to the case with stockouts and estimates demand and substitution rates using inventory-transactions data.

The assortment optimization problem (AP) is a nonlinear, nonseparable, discrete resource allocation problem. In §5, we describe an iterative heuristic that solves a series of separable nonlinear knapsack problems, and we assess its performance with a computational study. On average, the heuristic finds solutions within 0.5% of the optimal solution. The method presented in this paper is designed for large problems and can accommodate many realistic constraints (e.g., discrete maximum inventory levels, batch sizes, delivery lead times, and perishability of products).

The only known structural property of optimal assortments is related to demand rates (van Ryzin and Mahajan 1999). We establish properties of the heuristic solution that relate the products included in the assortment and their inventory levels to demand rates and other characteristics such as per-unit margin, per-unit volume, case-pack size, and demand variability.

We describe the details of the implementation at Albert Heijn, BV and assess its financial and strategical benefits in §6. Comparing the results of the recommendations of our system with the existing assortments suggests more than a 50% increase in profits.

The main contributions of this paper are the following. First, we are aware of no papers that provide empirical information about how assortment planning works in practice or evaluate a process against real data in the rapidly growing literature on assortment planning. We are seeing increasingly complicated models formulated and papers on algorithms or structural properties, with limited or no evidence as to the validity of those models, and not even a hint of how to estimate the parameters of these models. It seems timely for an injection of empirical evidence about the nature of real assortment planning problems. We provide a description of a real assortment planning problem together with real data, an approach for estimating parameters of the model, and a workable algorithm run on the real data to demonstrate its effectiveness. Second, we present a novel substitution estimation approach that works even when only sales summary data are available. The only paper on estimating substitution rates, Anupindi et al. (1998), requires inventory-transactions data and is limited to stockout-based substitution. For the cases when inventory-transactions data are available, we generalize the approach by Anupindi et al. (1998) to include a dynamic choice process by consumers. Third, we develop an iterative optimization heuristic for the assortment planning problem and establish new structural properties based on the heuristic solution.

2. Related Literature

For extensive reviews of the assortment planning literature, see Mahajan and van Ryzin (1998) and Kök et al. (2005). Below is a brief review of the related literature.

Assortment planning has been the focus of numerous industry studies, mostly concerned with the question of whether assortments were too broad or too narrow. Quelch

and Kenny (1994) report that the number of products in the marketplace increased by 16% per year between 1985 and 1992, while shelf space expanded by only 1.5% per year during the same period. Worried that current variety levels might be excessive, many retailers started 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).

Gruen et al. (2002) examine consumer response to stockouts across eight categories at 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% do not buy any item at all. Other studies of consumer response to stockouts have indicated that stockouts can entail substantial losses, both from a brand sales perspective (Schary and Christopher 1979) and from a category sales perspective (Fitzsimons 2000). 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 might be larger than those in case of an OOS, there are also significant similarities between consumer reactions in the two cases and, further, that OOS reactions for an item can be indicative of PAR responses for that item.

There are two common models of substitution. The utility-based model of substitution assumes that consumers associate a utility with each product in $N \cup \{0\}$, where $\{0\}$ denotes the no-purchase option, and that they choose the highest utility alternative available. The multinomial logit (MNL) model is a utility-based model that is commonly used in the economics and marketing literatures (e.g., Guadagni and Little 1983, Ben-Akiva and Lerman 1985, Anderson et al. 1992) and, more recently, in assortment planning models.

The MNL model assumes that the utility of alternative j to a particular customer has both a deterministic component u_j and a random component. The random component is assumed to follow a Gumbel (extreme value of Type I) distribution with mean zero and variance $\pi^2\mu^2/6$. The probability that alternative j is chosen is denoted p_j and is given by

$$p_j = \frac{e^{u_j/\mu}}{\sum_{k\in S} e^{u_k/\mu}}, \quad j \in S \cup \{0\}. \tag{1}$$

The MNL model in its simplest form is unable to capture an important characteristic of the substitution behavior because the rate of substitution is determined by the utility of the no-purchase option with respect to the utility of the products in S. 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 assumption 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 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 variant as a substitute, while the other half will prefer the no-purchase alternative to the other variant. 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-values. We can control the substitution rate by varying u_0, but that also determines the initial penetration rate of the category. Hence, with this model, it is not possible to have two categories with the same penetration rates but different substitution rates, which we have found severely limits the applicability of this model.

In the exogenous model of substitution, mostly used in inventory models (see Netessine and Rudi 2003 and the references therein), customers choose from the set N, and if the item they choose is not available for any reason, a customer might accept another variant as a substitute according to a given substitution probability. If the substitute is also unavailable, the sale is lost.

Van Ryzin and Mahajan (1999) study a stochastic single-period assortment planning problem under a multinomial logit choice model. In their model, consumers can substitute if their favorite variant is not carried (*assortment-based* substitution), but the sale is lost if their favorite variant is carried but temporarily unavailable (no *stockout-based* substitution). They show that the optimal assortment always consists of a certain number of the most popular products. This model is very stylistic because of restrictive assumptions such as identical costs, prices, and demand variability across products. Mahajan and van Ryzin (2001) develop a stochastic sample path optimization method for an assortment model with MNL choice and both types of substitution. Smith and Agrawal (2000) study the assortment planning problem with multiperiod base-stock inventory models under the exogenous model of substitution. They develop approximations of the objective function of the resulting integer program. They do not consider estimation of the model parameters.

Cachon et al. (2005) study the van Ryzin and Mahajan (1999) model in the presence of consumer search. Aydin and Hausman (2003) study assortment planning and supply chain coordination issues. Rajaram (2001) presents a mean-variance analysis of assortment planning with no substitution. Also related are multi-item inventory models, either with a single resource constraint (e.g., Nahmias and Schmidt 1984, Downs et al. 2001) or with substitutable products (e.g., McGillivray and Silver 1978, Parlar and Goyal 1984, Rajaram and Tang 2001, Avsar and Baykal-Gursoy 2002, Netessine and Rudi 2003). This group of papers does not consider the assortment problem and focuses on stocking decisions of the products in a given assortment.

The marketing and economics literatures have studied product variety extensively, mostly focusing on variety at the market level (e.g., Anderson et al. 1992, Shugan 1989) or from a product line design perspective (e.g., Green and

Krieger 1985, Moorthy 1984). We recognize that our model does not explicitly account for other possible factors that influence the relationship between assortment variety and demand: the space devoted to a category and the presence or absence of a favorite item influence the perception of variety (Kahn and Lehmann 1991, Broniarczyk et al. 1998), as do the arrangement, complexity, and presence of repeated items in an assortment (Hoch et al. 1999, Huffman and Kahn 1998, Simonson 1999).

Although assortment or inventory planning with substitutable products has attracted some attention in the literature, there is little existing work on the estimation of substitution behavior. Anupindi et al. (1998) describe a method for estimating consumer demand with stockout-based substitution (when the favorite variant is temporarily unavailable). They assume a Poisson arrival process for all products and find the maximum likelihood estimates of arrival rates and substitution probabilities via the expectation-maximization (EM) algorithm. Campo et al. (2003) also combine choice and availability data to measure the impact of stockouts in category sales. Talluri and van Ryzin (2004) utilize the EM algorithm to jointly estimate arrival rates and customer choice model parameters when no-purchase outcomes are unobservable. The estimation procedure in §4.3 is a generalization of Anupindi et al. (1998) and Talluri and van Ryzin (2004). The EM algorithm is first proposed by Dempster et al. (1977). Greene (1997) shows that the procedure converges under fairly weak conditions. Wu (1983) shows that the limiting value of the procedure would be a stationary point of the incomplete-data log-likelihood function if the expected log-likelihood function is continuous in the parameters.

In an influential paper, Corstjens and Doyle (1981) suggest a shelf-space allocation model by performing store experiments to estimate multiplicative sales and cost functions with own and cross-space elasticities. Their estimation and optimization procedures cannot be applied to large problems; as a result, they elected to work with product groups rather than SKUs. Neither Corstjens and Doyle (1981) nor the follow-up work explicitly model the assortment selection or the inventory side of the problem. Urban (1998) extends the basic model to include inventory-related costs and compare heuristic solutions.

Bretthauer and Shetty (2002) provide a review of nonlinear resource allocation models. Due to substitution, our objective function is a nonseparable function of the inventory levels. The quadratic knapsack problem is the only nonseparable problem that has been studied. Dussault et al. (1986) and Klastorin (1990) approximate the quadratic problem with a series of separable problems. Gallo et al. (1980) and Caprara et al. (1999) develop methods that directly solve the nonseparable problem with 0-1 variables.

3. The Substitution Model

Substitution behavior is not limited to the cases when consumers face stockouts. *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:

(1) The 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.

(2) The 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 might not carry that product. This is an example of assortment-based substitution.

(3) The 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 might be other products she might have preferred, but she didn't see them because either the retailer didn't carry them or they are stocked out. This could be an example for either substitution type, depending on whether the product is temporarily stocked out or not carried at that store.

The first two groups fit repeat purchases such as food, and the third fits one-time purchases such as apparel.

The substitution model considered in this paper is characterized by the following assumptions:

Assumption (A1). *Every customer chooses her favorite variant from the set N.*

Assumption (A2). *If for any reason this favorite is not available, 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 be specified differently for each round. Unlike the MNL model, this model can differentiate between categories that have the same initial demand for the category but different substitution rates through the choice of δ.

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

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

By limiting the number of substitution attempts, (A3) is not too restrictive. Kök (2003) shows that a multi-attempt model with δ' can be approximated with a single-attempt substitution model with rate $\delta'' > \delta'$, as long as δ' is not too large. In addition, a single-attempt model is similar to the utility-based MNL model, where the rate of substitution depends on the set of available products: Although δ is

fixed in our model, if a consumer cannot find her second-favorite product either, the sale is lost, and that is equivalent to less-frequent substitution when the set of available products is smaller.

The effective demand rate function under this substitution model is

$$D_j(\mathbf{f}, \mathbf{d}) = d_j + \left(\sum_{k \notin S} \alpha_{kj} d_k + \sum_{k \in S} \alpha_{kj} L_k(f_k, d_k) \right), \tag{2}$$

where the L_k function represents lost sales (average unmet demand) of product k. The first sum in (2) is the incremental demand for product j due to assortment-based substitution, and the second sum is the incremental demand for product j due to stockout-based substitution.

4. Demand Estimation

The data typically available for estimating the parameters of a demand model include the number of customers that made a transaction at each store on a given day, the sales for each product-store-day, and additional variables that influence demand, such as weather, holidays, and marketing variables (such as price and promotion). At Albert Heijn, assortment changes in permanent categories are done twice a year. Thus, at a particular store, all consumers see the same assortment (choice set) throughout the data collection period. At the time of our study, Albert Heijn was achieving 99.5% service level for nonperishable products, so it is reasonable to assume that a negligible level of stockout-based substitution was occurring. Therefore, we can estimate the demand of products carried in a store from store sales data. One possible estimation procedure is presented in §4.1, which is a variation of the methods widely used in the marketing literature and works well in our application. If the store carries less than a full assortment, then these demand estimates include the demand due to assortment-based substitution. We describe a procedure in §4.2 for estimating the assortment-based substitution rate using demand estimates from multiple stores, and we use this procedure to estimate the original demand from sales data that include assortment-based substitution. If the service levels are not high enough to ignore stockouts and stockout-based substitution but inventory-transactions data are available, one can estimate the original demand for each product and substitution probabilities (for both stockout- and assortment-based substitution) simultaneously by using a generalization of the demand estimation approach in §4.1. This generalized estimation approach is presented in §4.3.

4.1. Estimation of Demand for Products Carried in a Store

The estimation model in this section is based on the following assumption. The stockout rate is very low; therefore, it is reasonable to assume that stockouts and stockout-based substitution is negligible. We generalize the method to include stockouts and stockout-based substitution in §4.3.

Our model of consumer purchase behavior is based on three related 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*). This hierarchical model is standard in the marketing literature and is commonly used on panel data (e.g., Bucklin and Gupta 1992, Chintagunta 1993).

The demand for product j is

$$D_j = K(PQ)_j = K\pi p_j q_j, \quad j \in S, \tag{3}$$

where K is the number of customers who 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), 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 that a customer buys given purchase incidence and choice of product j.

Let the subscript h denote store, and t the day of the observation. In the grocery industry, the number of customers who visited store h on day t, K_{ht}, can be estimated by the daily number of customers who made transactions in that store. We use log-linear regression on transactions data to estimate κ_l, $l = 1, \ldots, 23$, in (4):

$$\ln(K_{ht}) = \kappa_1 + \kappa_2 T_t + \kappa_3 \mathrm{HDI}_t + \sum_{l=1}^{6} \kappa_{3+l} B_t^l + \sum_{l=1}^{14} \kappa_{9+l} E_t^l, \tag{4}$$

where the human discomfort index (HDI) is a combination of hours of sunshine and humidity, B^l is days of the week 0-1 dummies, and E^l is holiday 0-1 dummies for holidays, such as Christmas and Easter.

Purchase incidence is modeled as a binary choice:

$$\pi_{ht} = \frac{e^{v_{ht}}}{1 + e^{v_{ht}}}, \tag{5}$$

where the expected utility from the subcategory v is modeled as a linear function of various demand drivers for the subcategory, i.e., $v_{ht} = \gamma \nu_{ht}$.

We compute π_{ht}, the probability of purchase incidence for the subcategory, from sales data as the ratio of the number of customers who bought any product in S to the number of customers who visited store h on day t. We use the following logistic regression applied to sales history to estimate γ_l for $l = 1, \ldots, 24$ in (6):

$$\ln\left(\frac{\pi_{ht}}{1 - \pi_{ht}} \right) = \gamma \nu_{ht} = \gamma_1 + \gamma_2 T_t + \gamma_3 \mathrm{HDI}_t + \sum_{k=1}^{6} \gamma_{3+k} B_t^k + \gamma_{10} \bar{A}_{ht} + \sum_{l=1}^{14} \gamma_{10+l} E_t^l, \tag{6}$$

where T is the weather temperature and $\bar{A}$ is the average promotion level in the subcategory. $\bar{A} = \sum_j A_{jht} / |S|$, where

$A_{jht} = \{1$, if product j is on promotion on day t at store h; 0, otherwise$\}$. Other variables could be used as appropriate in a different context.

Product choice is modeled with the MNL framework, where p_{jht} is given by (1). The average utility of product j to a customer, u_{jht}, is assumed to be a function of product characteristics, marketing, and environmental variables, i.e., $u_{jht} = \beta w_{jht}$.

We compute p_{jht} from the sales data as the ratio of the number of customers who bought product j to the number of customers who 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 (1) (see Cooper and Nakanishi 1988 for details) to estimate β_l for $l = 1, \ldots, J+2$:

$$\ln\left(\frac{p_{jht}}{\bar{p}_{ht}}\right) = \beta w_{jht} = \sum_{k\in N} \beta_k I_{jk} + \beta_{J+1}(R_{jht} - \bar{R}_{ht}) + \beta_{J+2}(A_{jht} - \bar{A}_{ht}) \quad \text{for all } j \in S, \quad (7)$$

where $\bar{p}_{ht} = (\prod_{j\in S} p_{jht})^{1/|S|}$, $I_{jk} = \{1$, if $j = k$; 0 otherwise$\}$, R is price, and $\bar{R}$ is average price in the subcategory. It is straightforward to incorporate variables other than price and promotion into this approach.

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 ζ_l, $l = 1, \ldots, J+16$, in (8):

$$q_{jht} = \sum_{k\in N} \zeta_k I_{jk} + \zeta_{J+1} A_{jht} + \zeta_{J+2}\mathrm{HDI}_t + \sum_{l=1}^{14} \zeta_{J+2+l} E_t^l \quad \text{for all } j \in S. \quad (8)$$

We call this four-stage model the *choice-based approach*. The choice-based approach can be used in other contexts by including the relevant explanatory variables in (4)–(8). The current method used at Albert Heijn, called the *direct approach*, is estimating $(PQ)_j$ for each SKU directly via logistic regression, with similar explanatory variables. The quality of estimation is reported for both models in §6.2. The advantage of the choice-based model is that it imitates consumer choice behavior in a subcategory of substitutable products. In contrast, demand for SKUs in a subcategory is independent in the direct approach. Another advantage of the choice-based model, which will become apparent in §4.3, is that it can be generalized to estimate substitution behavior with inventory-transactions data, whereas it is not clear whether the direct approach is suitable for such a generalization. Nevertheless, this demand estimation module can be replaced with Albert Heijn's method or any other method without affecting the assortment-based substitution procedure and the optimization module. The only part of the paper that requires the choice-based approach is the stockout-based substitution estimation procedure in §4.3.

4.2. Estimation of Assortment-Based Substitution with Sales Summary Data

Suppose that a store carries assortment $S \subset N$ and that stockouts are negligible. We observe D_j for products $j \in S$ from sales data. Note that at a store with full assortment (i.e., $S = N$), no substitution takes place; hence $D_j = d_j$ for all j. If no store carries N, we redefine N as the broadest assortment carried in any store. Therefore, we can estimate d_j for $j \in N$ from the sales data of a similar store that carries a full assortment. If $\sum_{j\in S} D_j > \sum_{j\in S} d_j$, we conclude that the substitution rate in this subcategory is positive. Note also that our estimate of the underlying substitution rate must be higher if the gap is larger.

The substitution estimation method relies on comparing demand estimates from multiple stores. Rather than working with demand estimates, we use demand per customer, $(PQ)_j$, to eliminate the scale differences between stores. Unless it is necessary to do so explicitly, we omit the time subscript t for brevity.

After-Substitution Demand Estimation (ASDE) Models. We estimate the demand per customer via the application of the above models to data from a particular store h with assortment S_h.

$$(PQ)_{jh} = \pi p_{jh} q_{jh}, \quad j \in S_h \ \forall h$$

is the demand per customer for product j at store h, where the model (6)–(8) is calibrated using data from store h. Note that these demand estimates are for the effective demand, i.e., $D_j = K_h(PQ)_{jh}$, because they might include substitution demand if $N \backslash S_h \neq \varnothing$.

Original Demand Estimation (ODE) Model. No assortment-based substitution takes place at stores with full assortment. We estimate the original demand for the products by calibrating the above models with data from all full-assortment stores (i.e., h: $S_h = N$). In ODE, regression coefficients are common to all stores (i.e., the parameters are not store specific):

$$(PQ)^o_{jh} = \pi^o p^o_{jh} q^o_{jh}, \quad j \in N,\ h\text{: } S_h = N.$$

The superscript o denotes that the parameter estimates come from the ODE model.

After estimating the coefficients of the ODE model, we apply ODE to every store to estimate the original demand, i.e., $d_j = K_h(PQ)^o_{jh}$ for all h and for all $j \in N$. To account for store differences in the ODE models, we use additional explanatory variables in the regression equations (6)–(8). Albert Heijn considers *store size* (in square meters) and *percentage of customers who bike to the store* as defining store characteristics. Both have indications about the location of the store (urban-suburban) and the demographics of the customer base.

For full-assortment stores, both ODE and ASDE models estimate the original demand; however, ASDE models are store specific and therefore more accurate than ODE.

ODE models are used to estimate the original demand at all stores. Consider stores h' and h'', which have the same store properties, $S_{h'} = N$ and $N \backslash S_{h''} \neq \varnothing$. Then, $(PQ)^o_{jh'} = (PQ)^o_{jh''}$ for $j \in N$. That is, estimates of original demand d_j at a store with less than full assortment are the same as the demand of those products in a comparable store with full assortment.

Estimation of the Substitution Rate. For the purpose of estimating assortment-based substitution, we restrict the substitution matrix such that α_{kj} can be expressed as a function of a single subcategory substitution rate δ. We consider two models. Under random substitution, consumers choose their second-favorite product with equal probability:

$$\alpha_{kj} = \delta \frac{1}{|N|}, \quad k, j \in N. \tag{9}$$

Under proportional substitution, the rate of substitution to a product is proportional to the original demand rate of the product:

$$\alpha_{kj} = \delta \frac{d_j}{\sum_{l \in N \backslash \{k\}} d_l}, \quad k, j \in N. \tag{10}$$

These models reduce the number of substitution parameters to be estimated from J^2 to one. In both models, when product k is not available, $(1 - \delta)$ fraction of the demand for product k elects not to substitute, resulting in lost sales, and the δ fraction of demand is distributed to all other products. These one-parameter models capture many important real aspects of substitution. We are able to differentiate between subcategories with low and high substitution rates through the choice of δ. The following properties of the models are consistent with what would happen in a utility-based framework:

(i) $\alpha_{kj} \geqslant \alpha_{kl}$ if $d_j \geqslant d_l$, which implies that $D_j \geqslant D_l$ if and only if $d_j \geqslant d_l$.

(ii) Suppose that a store does not carry the full assortment, i.e., $N \backslash S \neq \varnothing$. Because only one round of substitution is allowed, the realized substitution rate from variant k to other products $\sum_{j \in S} \alpha_{kj}$ is increasing in set S. This means that a consumer who cannot find her favorite variant in the store is more likely to buy a substitute, as the set of potential substitutes gets larger.

(iii) Finally, in the proportional substitution matrix, the relative substitution rate from product k to i and from k to j is d_i/d_j. Recall that in the MNL model, that ratio is e^{u_i}/e^{u_j}, the ratio of the products' market shares.

Define the following auxiliary variables:

$y_h = \sum_{j \in S_h} (PQ)_{jh}$: subcategory total of after-substitution demand at store h given assortment S_h according to estimates from the ASDE model h.

$x_h = \sum_{j \in S_h} (PQ)^o_{jh}$: subcategory total of original demand at store h given assortment S_h according to estimates from the ODE model.

$z_h = \sum_{j \in N} (PQ)^o_{jh}$: subcategory total of original demand at store h given full assortment N according to estimates from the ODE model.

In words, y_h is the subcategory sales per customer at store h (according to the ASDE model), x_h is what store h would have sold with its current assortment (according to the ODE model) if there were no substitution, and z_h is what it would have sold (according to the ODE model) if it carried the full assortment. The most accurate estimate of observed sales per visiting customer at store h is y_h, but we do not know how much of y_h is substitution demand. For a given substitution rate δ, the effective demand for product $j \in S_h$ at store i according to the ODE model is

$$(PQ)^o_{jh} + \sum_{k \in N \backslash S_h} \alpha_{kj} (PQ)^o_{kh}. \tag{11}$$

Summing this over products $j \in S_h$ gives us the estimate $\hat{y}_h(\delta)$ according to the ODE model and the assumed substitution structure (i.e., what store h would have sold based on the ODE model estimates and substitution rate δ):

$$\hat{y}_h(\delta) = \sum_{j \in S_h} \left((PQ)^o_{jh} + \sum_{k \in N \backslash S_h} \alpha_{kj} (PQ)^o_{kh} \right).$$

Under random substitution, substituting α_{kj} from (9), we have

$$\begin{aligned} \hat{y}_h(\delta) &= \sum_{j \in S_h} (PQ)^o_{jh} + \sum_{j \in S_h} \sum_{k \in N \backslash S_h} \frac{\delta}{|N|} (PQ)^o_{kh} \\ &= x_h + \delta \frac{|S_h|}{|N|} (z_h - x_h). \end{aligned} \tag{12}$$

Similarly, under proportional substitution, substituting α_{kj} from (10), we have

$$\begin{aligned} \hat{y}_h(\delta) &= \sum_{j \in S_h} (PQ)^o_{jh} + \sum_{j \in S_h} \sum_{k \in N \backslash S_h} \delta \frac{K_h (PQ)^o_{jh}}{\sum_{l \in N - \{k\}} K_h (PQ)^n_{lh}} (PQ)^o_{kh} \\ &= x_h + \delta \sum_{j \in S_h} (PQ)^n_{jh} \sum_{k \in N \backslash S_h} \frac{(PQ)^o_{kh}}{\sum_{l \in N - \{k\}} (PQ)^o_{lh}} \\ &= x_h + \delta x_h \sum_{k \in N \backslash S_h} \frac{(PQ)^o_{kh}}{z_h - (PQ)^o_{kh}}. \end{aligned} \tag{13}$$

We choose the substitution rate to minimize the total squared error of estimation across stores h and time periods t:

$$\delta^* = \arg\min_{0 \leqslant \delta \leqslant 1} \sum_h \sum_t (\hat{y}_{ht}(\delta) - y_{ht})^2. \tag{14}$$

This procedure uses a combination of results from many regression models to estimate a substitution rate (i.e., purchase incidence, choice, and quantity models for each ASDE model and the ODE model). The total squared error in our predicted category sales (if we ignore substitution) is $\sum_h \sum_t (\hat{y}_{ht}(0) - y_{ht})^2$. Choosing δ^* to minimize this error gives $\sum_h \sum_t (\hat{y}_{ht}(\delta^*) - y_{ht})^2$. We measure percentage error reduction by comparing these two quantities. The substitution rate estimate becomes more significant if the percentage error reduction is higher. Results of the estimation at Albert Heijn are presented in §6.2.

Computation of the Original Demand Rates d_j. This involves two tasks: (i) deflating the demand rate of the variants already in the assortment S_h, and (ii) estimating a positive demand rate for the variants that are not in S_h. Clearly, if $S_h = N$, no computation is necessary.

Define T_h as the estimate of the total demand for products in N at store h. If the ODE model's predictions were perfect, T_h would be equal to $K_h z_h$. However, ODE models are less accurate than ASDE models because ASDE models are store specific. Hence, in estimating T_h, we scale $K_h z_h$ by the ratio of the ASDE model's estimate of total subcategory demand y_h to the total subcategory demand $\hat{y}_h(\delta^*)$ for S_h predicted by the ODE model:

$$T_h = K_h z_h \frac{y_h}{\hat{y}_h(\delta^*)}.$$

Based on the ODE models, x_h/z_h fraction of the total subcategory demand is for products in S_h and the remaining $(1 - x_h/z_h)$ is for products in N/S_h. We now allocate the total demand T_h between these two groups. The total demand for the products in S_h is $T_h x_h/z_h$. Because the ASDE model h has the most accurate information about relative sales of the products in S_h, we allocate this demand to individual products in S_h proportional to the demand estimates of the ASDE model:

$$d_{jh} = T_h \frac{x_h}{z_h} \frac{(PQ)_{jh}}{y_h} = K_h \frac{x_h}{\hat{y}_h(\delta^*)} (PQ)_{jh} \quad \text{if } j \in S_h. \tag{15}$$

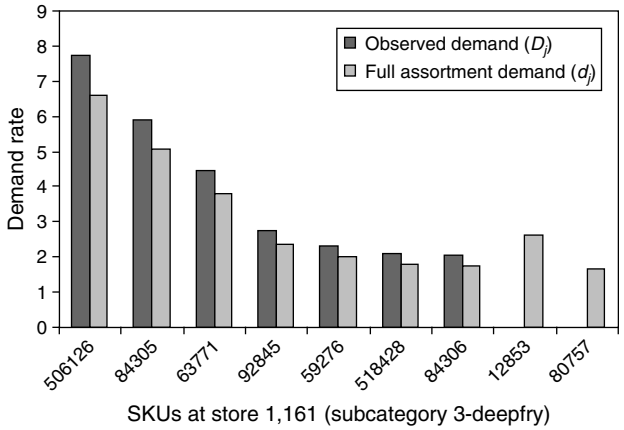
As one would expect, if the estimated substitution rate was zero, then $\hat{y}_h(\delta^*) = x_h$ due to Equations (12) and (13), and demand of the products in S_h remain as they were from the ASDE model, i.e., $d_{jh} = K_h (PQ)_{jh}$.

Total demand for the products that are not in S_h is $T_h(1 - x_h/z_h)$. Because store model h does not provide us with any information on how this demand should be allocated among products not in S_h, we allocate it proportionally to the demand rate estimates for store h of the ODE model (for example, if a small-sized product is selling better in stores with a higher percentage of customers who bike to the store, and h is such a store, we allocate a larger part of the total to that small-sized product):

$$\begin{aligned} d_{jh} &= T_h \left(1 - \frac{x_h}{z_h}\right) \frac{(PQ)^o_{jh}}{z_h - x_h} \\ &= K_h \frac{y_h}{\hat{y}_h(\delta^*)} (PQ)^o_{jh} \quad \text{if } j \notin S_h. \end{aligned} \tag{16}$$

Figure 1 presents an example of observed demand rates and the computed true demand rates for a subcategory with nine products. The right-most two products are not in S_h; therefore, their ASDE estimates are zero, but their true demand rate estimates are positive. The ratio of the demand for these two products to the total subcategory demand is the same as this ratio in the ODE models. The true demand rates of the first seven products (products in S_h) are lower than their ASDE estimates because the substitution demand is now removed.

Figure 1. Estimates of observed and original demand rates for a subcategory.



In our application, we estimate a substitution rate from assortment-based substitution and use it for both types of substitution. As Campo et al. (2004) points out, there are significant similarities in consumer reactions to a permanent assortment reduction and to stockouts. The advantage of estimating assortment-based substitution 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.

4.3. Estimation of Stockout-Based Substitution with Inventory-Transactions Data

We describe the algorithm for estimating stockout-based substitution using data from a single store. We later describe how the procedure can be applied to incorporate assortment-based substitution if data from multiple stores with nonidentical assortments are available. Because only a single store is considered, we drop the subscript h and assume that $S = N$.

Inventory-transactions data record the inventory levels of all products at every transaction such as replenishment or sales. From these data, we can infer the inventory levels of all products at all times. Sales transactions have records of all customers who visited the store and their times of departure. We assume that the number of customers who visited the store but did not purchase anything is negligible.

We first summarize the consumer behavior. Consider the customer who departs in time t. During her shopping, she elects to purchase from the subcategory with probability π_t given by (5). She then chooses her favorite product j with probability p_{jt} given by (1). Recall that π_t is a function of $v_t = \gamma \nu_t$ and p_{jt} is a function of $u_{jt} = \beta w_{jt}$. If the consumer's first choice is not available, then she substitutes according to substitution probability matrix $\alpha = (\alpha_{kj})_{k,\, j \in S}$.

Our objective is to obtain maximum likelihood estimates (MLE) of β, γ, and α. The general approach is to write

the likelihood function that represents the probability of observing the data, given our model of consumer behavior, and maximize it over the variables to be estimated. Even inventory-transactions data with this level of detail do not record every step of the above consumer behavior. Customers who are not interested in the subcategory, or who are originally interested but could not find an acceptable substitute, cannot be distinguished. Hence, inventory-transactions data are an incomplete data set. In what follows, we see that if complete data were available, we could write the likelihood function and maximize it relatively easily. However, in the presence of missing data, the function becomes complex and difficult to maximize. We deal with this missing-data problem by using the *expectation-maximization* (EM) algorithm. We first define the notation, derive the complete-data likelihood function and the incomplete-data likelihood function, and then describe the details of the EM algorithm. Define the following variables:

T: list of customer departure times indexed by t.
$S(t)$: set of available products at time t, $S(t) \subset S$.
$\bar{S}(t)$: set of products not available at time t, $\bar{S}(t) = S \backslash S(t)$.
$\bar{p}_{jt}$: probability that consumer t chooses option $j \in S \cup \{0\}$.

$$\bar{p}_{jt} = \begin{cases} \pi_t p_{jt} & \text{if } j \in S, \\ 1 - \pi_t & \text{if } j = 0. \end{cases}$$

$x(t)$: first choice, $x(t) \in S \cup \{0\}$.
$y(t)$: second choice (if applicable), $y(t) \in S \cup \{0\}$.
$z(t)$: observed choice, $z(t) \in S(t) \cup \{0\}$.

If we were to observe the customer choice behavior completely, the complete data set would be composed of the vectors $(S(t), x(t), y(t), z(t))_{t \in T}$. Note that $z(t) = x(t)$ if and only if $x(t) \in S(t) \cup \{0\}$. Also, $z(t) = y(t)$ if and only if $x(t) \in \bar{S}(t)$ and $y(t) \in S(t) \cup \{0\}$.

The data that can be inferred from the sales and inventory transactions data are $(S(t), z(t))_{t \in T}$. This data set is incomplete because $x(t)$ and $y(t)$ are not observed.

Define the following auxiliary variables for $j = 0, 1, \ldots, J$, $k = 0, 1, \ldots, J$, and $t \in T$:

$$X(t) = \{z(t)\} \cup \{j\colon j \in \bar{S}(t)\},$$

$$Y(j,t) = \begin{cases} \text{if } j = 0, \quad \text{then } \{z(t)\}, \\ \text{if } j > 0, \quad \text{then } \begin{cases} \{z(t)\} & \text{if } z(t) > 0, \\ X(t) - \{j\} & \text{if } z(t) = 0, \end{cases} \end{cases}$$

$$\theta(j,k,t) = \begin{cases} 1 & \text{if } x(t) = j \text{ and } y(t) = k, \\ 0 & \text{otherwise.} \end{cases}$$

$X(t)$ is the candidate set for the first-choice product. $Y(j,t)$ is the candidate set for the second choice given first choice j. $\theta(j,k,t)$ summarizes every choice made by a customer: if the customer at time t purchased her first choice j, then $\theta(j,j,t) = 1$. If a customer could not find her first choice j and purchased her second choice k, then $\theta(j,k,t) = 1$ for $k \neq j$ ($k = 0$ means she purchased "nothing"). Note that $\theta(j,k,t)$ can take positive values only for $j \in X(t)$ and $k \in Y(j,t)$. The missing data are $\theta(j,k,t)$ for $j \in X(t)$, $y \in Y(j,t)$, $t \in T$.

The likelihood function based on the incomplete data set $(S(t), z(t))_{t \in T}$ is

$$\prod_t \left(\bar{p}_{z(t),t} + \sum_{k \in \bar{S}(t)} \alpha_{k,z(t)} \bar{p}_{kt} \right).$$

Because the logarithm is an increasing function, maximizing the log-likelihood function and the likelihood function are equivalent. Anupindi et al. (1998) shows that the log-likelihood function of a model equivalent to the above model is concave in $\bar{p}_j$ when $\bar{p}_j$ is exogenous and stationary (i.e., $\bar{p}_{jt} = \bar{p}_j$ for all t). However, it is not clear whether the likelihood function is concave in (γ, β, α), and the numerical optimization of this function is difficult.

An alternative approach that makes it easier to find the MLE for this problem is the *expectation-maximization* (EM) algorithm. The EM algorithm is the most widely used statistical method for missing data problems. It uses the log-likelihood function based on the complete data set in an iterative algorithm starting with arbitrary (γ, β, α). The E-step replaces the incomplete data with their expectation using the current estimates. The M-step maximizes the complete-data likelihood function to obtain new estimates. The procedure is repeated until the parameter estimates converge. Because we cannot establish concavity, it is possible that the EM algorithm converges to a local optimum. One way to deal with this problem is to try different starting points. The advantage of the procedure is that maximizing the complete-data likelihood function is much easier than maximizing an incomplete-data likelihood function.

Fix $\alpha_{jj} = 1$ for all $j = 0, 1, \ldots, J$ and $\alpha_{0j} = 0$ for all $j = 1, \ldots, J$. The complete-data likelihood function is

$$\mathrm{L}(\gamma, \beta, \alpha) = \prod_t \prod_{j \in X(t)} \prod_{k \in Y(j,t)} (\bar{p}_{jt} \alpha_{jk})^{\theta(j,k,t)}.$$

Define $\Theta(j,t) = \sum_{k \in Y(j,t)} \theta(j,k,t)$. The log-likelihood function is

$$\begin{aligned} \mathcal{L}(\gamma, \beta, \alpha) &= \sum_t \sum_{j \in X(t)} \sum_{k \in Y(j,t)} \theta(j,k,t)(\ln \bar{p}_{jt} + \ln \alpha_{jk}) \\ &= \sum_t \sum_{j \in X(t) \backslash \{0\}} \Theta(j,t) \ln \pi_t + \sum_t \sum_{0 \in X(t)} \Theta(j,t) \ln(1 - \pi_t) \\ &\quad + \sum_t \sum_{j \in X(t) \backslash \{0\}} \Theta(j,t) \ln p_{jt} \\ &\quad + \sum_t \sum_{j \in X(t)} \sum_{k \in Y(j,t)} \theta(j,k,t) \ln \alpha_{jk}. \end{aligned} \tag{17}$$

Note that the log-likelihood function is composed of three separate parts with π, p, and α. Hence, it is separable in γ, β, and α, and the maximization of each part can be done separately. The part with π_t can be rewritten as

$$\sum_t \sum_{j\in X(t)\setminus\{0\}} \Theta(j,t)(\gamma\nu_t - \ln(1+e^{\gamma\nu_t})) - \sum_t \sum_{0\in X(t)} \Theta(j,t)\ln(1+e^{\gamma\nu_t})$$
$$= \sum_t \sum_{j\in X(t)\setminus\{0\}} \Theta(j,t)\gamma\nu_t - \sum_t \sum_{j\in X(t)} \Theta(j,t)\ln(1+e^{\gamma\nu_t}). \quad (18)$$

This is identical to the log-likelihood function of the standard binary choice model, which is jointly concave in γ (Greene 1997). The part of $\mathcal{L}$ with p_{jt} can be rewritten as

$$\sum_t \sum_{j\in X(t)\setminus\{0\}} \Theta(j,t)\left(\beta w_{jt} - \ln\sum_{l\in S} e^{\beta w_{lt}}\right). \quad (19)$$

This is identical to the log-likelihood function of the standard MNL model, which is jointly concave in β under fairly general conditions (McFadden 1974). Finally, the part with α is also concave. As a result, with complete data we can obtain the MLE estimates for all parameters easily.

Each iteration of the EM algorithm consists of two steps.

Step 1. The E-step: We compute the expected value of the missing data given (γ, β, α).

Let $\hat{\theta}(j,k,t)$ denote the expectation of $\theta(j,k,t)$ for $j \in X(t), k \in Y(j,t), t \in T$.

$$\hat{\theta}(j,k,t)$$
$$= E[\theta(j,k,t) \mid z(t), (\gamma,\beta,\alpha)]$$
$$= \Pr\{\theta(j,t) = 1 \mid z(t), (\gamma,\beta,\alpha)\}$$
$$= \frac{\Pr\{z(t) \mid \theta(j,t)=1, (\gamma,\beta,\alpha)\}\Pr\{\theta(j,k,t)=1, (\gamma,\beta,\alpha)\}}{\Pr\{z(t) \mid (\gamma,\beta,\alpha)\}}$$
$$= \frac{\bar{p}_{jt}\alpha_{jk}}{\sum_{i\in X(t)}\sum_{l\in Y(i,t)}\bar{p}_{it}\alpha_{il}}.$$

$\hat{\theta}(j,k,t)$ is the conditional probability that the customer in time t choose first j then k given $z(t)$ and $S(t)$.

Step 2. The M-step: Given the estimates of the missing data, we replace $\hat{\theta}(j,k,t)$ for $j \in X(t)$, $k \in Y(j,t)$, $t \in T$ in the complete data set to obtain the expectation of the log-likelihood function. This is possible because the log-likelihood function is linear in the missing data.

The maximization of (18) with respect to γ and the maximization of (19) with respect to β can be done via standard nonlinear optimization techniques. If $(\bar{p}_0, \bar{p}_1, \ldots, \bar{p}_J)$ was exogenous rather than the outcome of a choice model, we could obtain closed-form expressions for the MLEs. To maximize (17) subject to $\sum_{j\in S\cup\{0\}} \bar{p}_j = 1$, we write the Lagrangian relaxation with dual variable η_0:

$$\sum_t \sum_{j\in X(t)} \Theta(j,t)\ln\bar{p}_j - \eta_0\left(\sum_j \bar{p}_j - 1\right).$$

The solution to the first-order condition for $\bar{p}_j$ yields

$$\bar{p}_j = \frac{1}{\eta_0}\sum_{t:\, j\in X(t)} \Theta(j,t).$$

Substitute these in $\sum_{j\in S\cup\{0\}} \bar{p}_j = 1$ and solve to obtain the unique MLE estimates:

$$\bar{p}_j = \frac{\sum_{t:\, j\in X(t)}\Theta(j,t)}{\sum_{i\in S\cup\{0\}}\sum_{t:\, i\in X(t)}\Theta(i,t)} = \frac{1}{|T|}\sum_{t:\, j\in X(t)}\Theta(j,t).$$

The estimate for the probability of j being the first choice is simply the number of observations with j as the first choice divided by the total number of customers.

Finally, to obtain the estimate of α, we maximize (17) subject to the following constraints:

$$\sum_{k\in S\setminus\{j\}} \alpha_{jk} = 1 \quad \text{for all } j \in S. \quad (20)$$

The Lagrangian relaxation with dual variables η_j is

$$\sum_t\left[\sum_{j\in X(t)}\sum_{k\in Y(j,t)}\theta(j,k,t)\ln\alpha_{jk}\right] - \sum_{j\in S}\eta_j\left(\sum_{k\in S\setminus\{j\}}\alpha_{jk} - 1\right).$$

The solution to the first-order condition for α_{jk} yields

$$\alpha_{jk} = \frac{1}{\eta_j}\sum_{t:\, j\in X(t),\, k\in Y(j,t)}\theta(j,k,t), \quad j,k \in S,\ j \neq k. \quad (21)$$

Replacing α_{jk} in (20) and solving for η_j, we get

$$\eta_j = \sum_{l\in S\setminus\{j\}}\sum_{t:\, j\in X(t),\, l\in Y(j,t)}\theta(j,l,t).$$

Replacing that in (21), we obtain the estimates of the substitution probability matrix:

$$\alpha_{jk} = \frac{\sum_{t:\, j\in X(t),\, k\in Y(j,t)}\theta(j,k,t)}{\sum_{l\in S\setminus\{j\}}\sum_{t:\, j\in X(t),\, l\in Y(j,t)}\theta(j,l,t)}, \quad j,k \in S,\ j \neq k.$$

This is also quite intuitive. The probability of substitution from j to k is the number of observations with j as the first choice and k as the second choice, divided by the number of observations with j as the first choice.

The arrival rate to the store λ can be estimated by dividing the total number of customers that visited the store on a given day (K) by the length of the day. A more refined approach would be to consider a nonstationary arrival process and estimate $\lambda(t)$ for different times of the day.

The estimation procedure above is a generalization of the approaches in Anupindi et al. (1998) and Talluri and van Ryzin (2004). Anupindi et al. (1998) estimate stationary λ, $\bar{p}$, and α, but do not consider the dynamic choice process—that is, the dependence of $\bar{p}$ on time and other variables through γ and β. 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. Our approach combines these approaches and estimate $(\lambda, \gamma, \beta, \alpha)$, where the consumers' original choice is based on the MNL model (and the values of the marketing and other variables) and stockout-based substitution is governed by a general probability matrix.

This procedure can be extended to multiple stores by simply including the store index h in all variables and all sums. Any of γ, β, and α might or might not be store specific. If they are all store specific, however, then each store's estimation problem is independent of the others. Let (γ, β, α) be common across stores. The procedure can be extended to include assortment-based substitution by considering $S_h = N$ for all h. Of course, the products that are not carried in a store would always be in $\bar{S}_h(t)$. The procedure can also be modified to estimate different probability matrices α^s for stockout-based substitution and α^a for assortment-based substitution. This modification would require redefinition of the auxiliary variables and sets. It is also necessary that the assortments across stores sufficiently vary to be able to estimate all elements of α^a.

5. Assortment Optimization

Recall the objective function of the optimization problem (AP):

$$Z(\mathbf{f}) = \sum_j G_j(f_j, D_j(\mathbf{f}, \mathbf{d})).$$

In the inventory system described in §1, 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.

We propose the following iterative heuristic that solves a series of separable problems. 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 Equation (2) and then solve (AP). We keep iterating until f_j^t converges for all j.

Iterative Heuristic (IH)

Step 1. Set $t = 0$. Solve (AP) with $Z(\mathbf{f}) = \sum_j G_j(f_j, d_j)$ via a greedy heuristic and record the solution as $\mathbf{f}^0$.

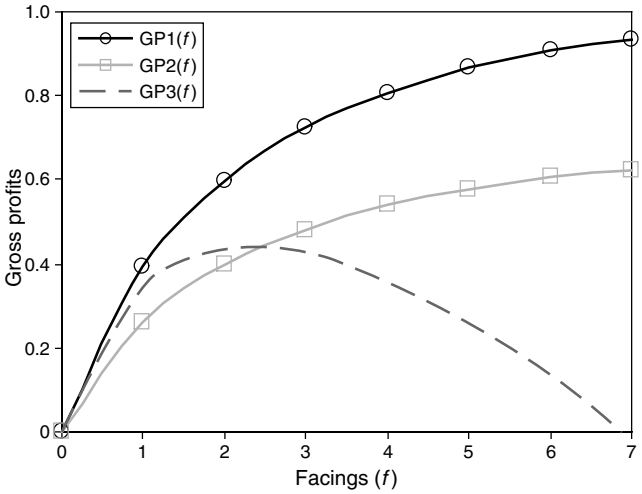
Step 2. $t \leftarrow t + 1$.

Step 3. Solve (AP) with $Z(\mathbf{f}^t) = \sum_j G_j(f_j^t, D_j(\mathbf{f}^{t-1}, \mathbf{d}))$ via a greedy heuristic and record the solution as $\mathbf{f}^t$.

Step 4. If $f_j^t \neq f_j^{t-1}$ for all j, then the facing allocation has changed, GO TO 2. If not, then the procedure converged, with solution $\mathbf{f}^t$.

We estimate $G_j(f_j, D_j)$ by simulating the replenishment of product j in isolation from other products. In Step 1 of (IH), we simulate each product with demand rate d_j and record $G_j(f_j, d_j)$ and $L_j(f_j, d_j)$ for $f_j = 1, 2, \ldots$. In Step 3, we simulate each product with demand rate D_j, where lost sales estimates in the computation of D_j via (2) are those recorded in Step 1. Figure 2 illustrates the gross profit curves for three SKUs. The details of the simulations are discussed in §6.

Figure 2. Simulation estimates of gross profit curves for three SKUs (1 and 2 are nonperishable and 3 is perishable).



Step 3 of the algorithm specifies an objective function that is separable in the allocated facings (i.e., G_j depends only on f_j^t). Hence, (AP) becomes a separable knapsack problem. The optimization in Step 3 is done by a greedy heuristic.

Because Step 3 assumes independent products, the solution $\mathbf{f}^t$ may imply a demand vector different than the inputted demand vector. (IH) has converged when there is consistency between the input demand vectors and the resulting solution. We do not know the true value of the objective function until (IH) converges. Therefore, the objective function value does not necessarily improve at each iteration. The convergence of the algorithm is not guaranteed. However, it never failed to converge in our numerical tests. In the application, the algorithm stops when no further improvement in the objective function value is achieved for a number of iterations or when a maximum number of iterations is reached.

(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 the subscript $i = 1, \ldots, I$ 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.

We propose the addition of the following local search to the (IH) algorithm. The local search consists of making small perturbations to the current solution and observing the final impact of that perturbation.

Local Search

Step 5. $\mathbf{f}^* \leftarrow (\mathbf{f}_i^t)_{i=1,\ldots,I}$.

Step 6. Randomly pick a subcategory i that was not picked before in local search and that has positive variety, i.e., $\sum_j f_{ij}^* > 0$. If none left, STOP.

Step 7. Drop the last product in subcategory i that is added to the assortment and mark that product to not enter the assortment again.

Step 8. Apply Steps 2 through 4. When the procedure converges, if $Z(\mathbf{f}) > Z(\mathbf{f}^*)$, then $\mathbf{f}^* \leftarrow (\mathbf{f}_i^t)_{i=1,\ldots,I}$; otherwise, $(\mathbf{f}_i^t)_{i=1,\ldots,I} \leftarrow \mathbf{f}^*$.

Step 9. GO TO Step 6.

In the case of multiple subcategories in (AP), the facing allocations for SKUs also determine a shelf-space allocation between subcategories. The shelf-space allocation between categories for the whole store can even be determined in this way. This can be viewed as a bottom-up approach to allocating shelf space in a store, as opposed to a top-down approach, where store space is allocated to categories, then category space is allocated to subcategories, and finally subcategory space to SKUs. In our application, the two approaches are blended together: category shelf space is determined by an external algorithm that is used by Albert Heijn. Then, the space for each subcategory is determined as a result of the facing allocation to SKUs in (AP) given the category shelf space.

For nonseparable resource allocation problems, another iterative method guaranteed to reach the optimal solution in quadratic optimization problems is found in Dussault et al. (1986) and Klastorin (1990). They simply solve the integer problem with the branch-and-bound method, but at each node of the branch-and-bound, method, solve the relaxed (i.e., continuous variables) nonseparable quadratic optimization problem with a series of separable problems. At each iteration, Taylor approximation is updated based on the current solution. Our method applies a similar idea to the discrete space. The application of the greedy heuristic in Step 3 corresponds to the solution of the approximate separable objective function, and the updating of the demand vector using the current solution corresponds to the recomputation of the Taylor approximation. The difference is that our method incorporates integrality at each iteration, whereas Dussault et al. (1986) and Klastorin (1990) impose integrality constraints at the highest level of hierarchy in the branch-and-bound method. Another difference is that the quadratic problem's Taylor approximation is clearly separable, whereas because it is hard to characterize the shape of the dependence at the objective function level, we impose separation at the input level (i.e., the demand vector) rather than the objective function.

5.1. Numerical Study of the Iterative Heuristic

This section assesses the performance of the iterative heuristic. We study three sets of examples. The first set includes 14 single-subcategory problems with up to seven SKUs in each subcategory. The second set includes 13 two-subcategory problems with up to a total of eight SKUs between the two subcategories. The third set includes four three-subcategory problems with up to a total of eight SKUs between them. The subcategories, information on SKUs, and demand estimates are taken from the Albert Heijn application. We consider the combination of the following settings resulting in six instances for each of the 31 problems:

$$\delta = \{0, 1\},$$

$$\textit{shelf space} = \{1/6, 1/3, 2/3\} * (\textit{full\_service\_shelf\_space}),$$

where *full_service_shelf_space* is set such that we have sufficient space to carry all products with a service level of 99.9% or more. We solve for the optimal solution via enumeration and compare the heuristic solution to the optimal gross profits. To reduce running time, we ignore the characteristics of the products and of the replenishment system and use a newsvendor profit function to compute the gross profit, i.e., $G_j(f_j, D_j) = m_j \mathrm{E}[\min(D_j, c_j f_j)]$ and $L_j(f_j, d_j) = \mathrm{E}[\max(0, d_j - c_j f_j)]$.

Table 1. Performance of the iterative heuristic in comparison to the optimal solution.

	$\delta_i = 0$ for all i	$\delta_i = 1$ for all i	
	Iteration 1	Iteration 1	After convergence
Average % gap	0.1	3.0	0.5
Maximum % gap	3.2	24.1	12.1
Instances with zero-gap out of 93 instances	85	14	40

Table 1 summarizes the results. In the $\delta = 0$ case, the initial greedy solution (Iteration 1 of the iterative heuristic) is the final solution of the heuristic. The average percentage gap with the optimal solution is 0.1% with a maximum of 3.2%. Iteration 1 of the iterative heuristic found the optimal solution 85 times out of 93.

In the $\delta = 1$ case, the result from Iteration 1 has an average gap of 3.0% and a maximum gap of 24.1%, finding the optimal solution in 14 times out of 93. The iterative heuristic reduces the average gap to 0.5%, the maximum gap to 12.1%, and finds the optimal solution in 40 instances (26 more instances than in the first iteration). The 2.5% difference on average between the iterative heuristic solution and the solution from the first iteration is simply the value of explicitly considering substitution in optimization. Table 2 shows that the performance of the heuristic is very good, particularly with high levels of shelf space.

5.2. Structural Properties of the Iterative Heuristic

In this section, we characterize the properties of the resulting assortment from the iterative heuristic. We describe the impact of product characteristics such as gross mar-

Table 2. Performance of the optimization method in comparison to the optimal solution for different levels of shelf space.

	Shelf space (%)		
	Low	Medium	High
Iteration 1			
Average % gap	3.2	1.5	0.0
Maximum % gap	24.1	12.1	0.4
After convergence			
Average % gap	0.5	0.4	0.0
Maximum % gap	12.1	6.5	0.1

gin, space requirement, coefficient of variation of demand, and case sizes on product selection and product inventory levels. We start by defining the characteristics of the products and the replenishment system.

Products A and B belong to a subcategory with substitution rate $\delta \geqslant 0$. They are nonperishable. They are subject to the replenishment system described in §1. The lead time is zero. Demand for both products follows 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, $d_A = d_B$, $\rho_A = \rho_B$, $c_A = c_B$, and $b_A = b_B = 1$.

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 sizes and delivery lead time, affect the curvature of the gross profit function. A product with a smaller case-pack (batch size) has a higher slope of the gross profit curve for low inventory levels; therefore, it can achieve the maximum gross profit with lower inventory levels. The following lemmas state these observations formally.

LEMMA 1. *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* $d_A > d_B$. *The demand distribution is one of Poisson, exponential or normal distribution.* α *satisfies the condition that* $D_j \geqslant D_k$ *if and only if* $d_j \geqslant d_k$.

(ii) *All else is equal and per-unit gross margins* (m) *of A and B are such that* $m_A > m_B$. *Then,*

$$\Delta G_A(f, D_A) > \Delta G_B(f, D_B) \quad \text{for } f \geqslant 1 \text{ and any } \tilde{\mathbf{f}},$$

where $\Delta G(f, D) \equiv G(f, D) - G(f-1, D)$.

PROOF. First note that because D_A is fixed for given $\tilde{\mathbf{f}}$ and f_B, $G_A(\,)$ is a function of f_A only.

(i) Because the substitution demand is allocated to the products proportional to the original demand rates, the mean of effective demands for products A and B follow the order of the mean of original demands, i.e., $d_A > d_B \Rightarrow D_A > D_B$ for any $\tilde{\mathbf{f}}$. Therefore, we can compare the gross profit functions based on the original demand rates.

Let $\tilde{D}_A$ and $\tilde{D}_B$ denote the random demand variables with mean D_A and D_B, respectively. The inventory level after each order is $c_A f_A$ and $c_B f_B$ for A and B, respectively. At each period, expected sales are $\mathrm{E}[\min(\tilde{D}_A, c_A f_A)]$ and $\mathrm{E}[\min(\tilde{D}_B, c_B f_B)]$. Clearly, increasing the demand rate increases sales, i.e., $G_A(f, D_A) > G_B(f, D_B)$. The marginal benefit of increasing the maximum inventory level evaluated at inventory level I is $\Pr\{\tilde{D}_A \geqslant I\}$. Because $D_A > D_B$, it is easy to show that $\tilde{D}_A$ is stochastically larger than $\tilde{D}_B$ for the three distributions listed, i.e., $\Pr\{\tilde{D}_A \geqslant I\} \geqslant \Pr\{\tilde{D}_B \geqslant I\}$ for all I. Hence, $\Delta G_A(f, D_A) > \Delta G_B(f, D_B)$.

(ii) Because the substitution demand is allocated to the products that are in the assortment proportional to the original demand rates, $D_A = D_B$ for any $\tilde{\mathbf{f}}$. G is sales times per-unit gross margin. Because expected sales is the same for A and B and $m_A > m_B$, we have $\Delta G_A(f, D_A) = (m_A/m_B)\Delta G_B(f, D_B) > \Delta G_B(f, D_B)$. □

The condition on the substitution matrix in part (i) of Lemma 1 is reasonable for a homogenous set of products, consistent with substitution under the MNL model, and it is satisfied by the random and proportional substitution matrices in (9) and (10). Note that Lemma 1 also implies that $G_A(f, D_A) > G_B(f, D_B)$ for any f under both cases (i) or (ii). We suspect that similar results to these lemmas hold for replenishment systems with continuous time, continuous or periodic review, and positive lead times.

LEMMA 2. *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* $\rho_A < \rho_B$,

(ii) *All else is equal,* $b_A \geqslant 1$, *and* b_B *is an integer multiple of* b_A,

then, the following holds for any $\tilde{\mathbf{f}}$:

$$G_A(f, D_A) > G_B(f, D_B) \quad \text{for } f \geqslant 1$$

and

$$\lim_{f\to\infty} G_A(f, D_A) = \lim_{f\to\infty} G_B(f, D_B).$$

PROOF. The second result is obvious: given infinite inventory levels, sales for both products are equal to the expected effective demand rates and we have $D_A = D_B$ for any $\tilde{\mathbf{f}}$ because $d_A = d_B$.

(i) At every period, the initial inventory level is the same for both products. Note that the mean effective demand for products A and B is the same. Expected sales is mean demand − lost sales during each period. Lost sales in a period increases with demand variance. Hence, sales of product A is larger than product B.

(ii) Product A's replenishment system can always replicate product B's ordering decisions. However, because it is more flexible, at any ordering point, A orders more than or equal to the order size of B, resulting in more inventory for A. This means lower lost sales and higher gross profits in each period. □

The following theorems characterize properties of the iterative heuristic solution based on the above lemmas and the properties of the greedy heuristic.

THEOREM 3. *Consider products A and B. If exactly one of the following conditions of Lemma* 1 *or condition* $w_A \leqslant w_B$ *is met, then* $f_A \geqslant f_B$ *in the final solution of the iterative heuristic.*

PROOF. Recall that w_j is the width of a facing of product j. In the greedy allocation step of each iteration, the value of assigning the fth facing to product j per-unit shelf space is $\Delta G_j(f, D_j)/w_j$. In parts (i) and (ii), the value of the nominator is higher for product A than B by Lemma 1. In part (iii), the value of the denominator is smaller. Hence, the greedy heuristic will never assign facing f to product B before product A. □

If the first condition holds, the implications of this proposition 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). When the second (third) condition holds, the proposition implies that more inventory should be allocated to products with higher gross margin (lower space requirement).

THEOREM 4. *Consider products A and B. If exactly one of the conditions of Lemma* 2 *is met, 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.,* $f_B > 0 \Longrightarrow f_A > 0$).

PROOF. In the greedy allocation step at each iteration, the value of assigning the first facing to product j per-unit shelf space is $GP_j(1, D_j)/w_j$. By Lemma 2, the marginal value is higher for product A than B. Hence, the greedy heuristic will never assign the first facing to product B before product A. □

When one of the conditions of Lemma 2 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, then neither is B. Because 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 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.

To summarize, products with higher demand, higher margin, or smaller physical size should be included first in the assortment and should be assigned more inventory. Products with lower demand variability and smaller case sizes should also be included in the assortment first, but more inventory can be assigned to products with higher demand variance and larger case sizes if the available shelf space is sufficiently high.

6. Application

6.1. Data

At Albert Heijn, the data set included SKU-day-store level sales data through a period of 20 weeks in seven merchandise categories from 37 Albert Heijn stores. For each store day, we know the number of customers who made a transaction in the store. For each SKU-day-store, we know sales data including the number of units sold, the number of customers who purchased that product, selling price, and whether or not the product was on promotion. In addition, we have daily weather data and a calendar of holidays such as Christmas and Easter. The categories are cereals, bread spreads, butter & margarine, canned fruits, canned vegetables, cookies, and banquet sweets. There were 114 subcategories and 880 SKUs in these seven categories. The number of subcategories and SKUs in each category is shown in Table 3. The size of subcategories varies from 1 to 29 SKUs, with an average of 7.7 and a standard deviation of 5.7.

6.2. Estimation Results

Demand Estimation. The quality of the estimation by models (6)–(8) is tested by reporting the ratio of mean absolute deviation to average demand for each product. Data from the first 16 weeks are used to calibrate (fit) the regression models, and data from the last four weeks are used to test the prediction ability.

We first report the results for the ASDE models. Table 4 reports average mean absolute deviation (MAD) figures for both fit and test samples for seven merchandise categories.

Table 3. Summary of merchandise categories.

Category name	# Subcategories	# SKUs
1	24	219
2	18	58
3	25	177
17	16	260
53	14	55
55	9	55
127	8	56
Total	114	880

Table 4. Comparison of the demand estimation models for products in the store.

	ASDE models (%)				ODE models (%)				SKU-level logistic models (%)			
	MAD		Bias		MAD		Bias		MAD		Bias	
Categories	Fit	Test	Fit	Test	Fit	Test	Fit	Test	Fit	Test	Fit	Test
1	76	81	−7	−5	88	92	−6	−1	74	81	−51	−53
2	90	80	−1	−6	81	90	−14	−5	75	89	−53	−47
3	76	81	−7	−5	86	91	−8	−3	75	81	−54	−53
17	77	79	−6	−2	99	98	6	10	74	117	−44	−17
53	58	58	−5	−4	77	75	6	7	62	68	−34	−39
55	39	42	−4	−3	51	56	0	5	47	51	−24	−22
127	77	79	−9	−7	93	93	−6	−3	76	88	−55	−49
Grand avg.	74	76	−6	−4	87	90	−3	2	72	89	−47	−40

The purchase incidence model has an average MAD of 22%. MAD of the choice probabilities range from 15% for fast movers to 100% for slow movers, with an average MAD of 40% across all products. MAD of quantity regression is 15% on average. The average MAD of $(PQ)_j$ across all products, subcategories, and stores is 74% in the fit sample and 76% in the test sample. The average bias of our approach is −6% and −4% in the test and fit samples, respectively. The average MAD for the ODE models across all SKUs and stores is 84% and 86% in the fit and test samples, respectively.

We compare the quality of estimation by the ASDE models with the current estimation method at Albert Heijn (SKU-level logistic regression). Albert Heijn's approach yields a MAD of 72% and 89% and an average bias of −47% and −40% in the fit and test samples, respectively. Therefore, we conclude that our approach slightly outperforms the current approach in the quality of the estimation.

Substitution Rate. Among the 114 categories, 48 have full assortment at all 37 stores. Therefore, we were able to estimate the substitution rate for only the remaining 66 subcategories, for which the estimation results and the percentage error reduction are presented in Table 5 under random and proportional substitution. In both cases, 32 subcategories out of 66 have positive substitution rate estimates, and average error reduction among those is 14% for proportional and 10% for random substitution. It is important to note that the estimates of δ under random and proportional substitution are either the same or very close for most subcategories. In the very few cases that the models' substitution rate estimates disagree, the error reduction is less than 1%. We conclude that we can measure the rate of substitution in a subcategory with either substitution model, although the proportional substitution model seems to perform slightly better.

6.3. Application Details

As mentioned before, we use simulation to estimate each product's average gross profit for the given number of facings (in isolation from other products) because we are not able to obtain closed-form or approximate expressions for the complicated inventory model used at Albert Heijn. In previous studies, Albert Heijn established that demand follows a Poisson distribution for slow-moving products (i.e., product with a daily demand rate less than 10 units) and gamma distribution with standard deviation equal to the square root of the mean for fast-moving products (i.e., daily demand rate higher than 10 units). For perishable products, the age of individual units is tracked in the simulation, and products reaching their shelf life are disposed. The simulation results are averages from multiple replications. In each replication, the first 10 periods are the warm-up periods, and statistics are recorded for the last 250 periods. We continue to replicate until a 95% confidence interval for the average gross profit is reached.

There are other operational constraints that are incorporated into the optimization module for the real application. The minimum number of SKUs and the minimum number of facings in a subcategory and the minimum and maximum number of facings for particular SKUs are specified by merchandising managers. It is not difficult to incorporate these constraints into the solution of Steps 1 and 3 of the iterative heuristic. The merchandising managers also have the option to override the recommended substitution rates.

The delivery schedule for each category of products is known. A period is defined as the time between two deliveries. We estimate the original demand for each SKU for each period by using the methodology in §4.2. For each subcategory, we choose the period that has the highest total demand across SKUs. We call this the *peak-load period*. The optimization is done for the peak-load period demands. For nonperishable items, the assigned facings are filled as much as possible (following inventory policy) at all times, even during nonpeak-load periods. For perishable items such as produce that have a shelf life of a few days or less, the recommended facing allocations determine which products will be carried in a store, but inventories are controlled dynamically: 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

Table 5. Substitution estimates from different weeks in the planning period.

						Proportional substitution		Random substitution		
Group	Subcategory	$\|N\|$	Min$\|S\|$	Number of stores with $\|S\| < \|N\|$	Total error ($\delta = 0$)	δ^*	% Error reduction[a] (%)	δ^*	% Error reduction[a] (%)	Models disagree
1	3	14	13	6	78.2	0	0.0	0	0.0	
1	5	8	5	7	6.2	0	0.0	0	0.0	
1	13	4	2	10	6.6	1	0.5	1	1.1	
1	19	23	20	11	135.0	0	0.0	0	0.0	
1	20	10	9	5	54.7	0	0.0	0	0.0	
1	25	7	6	4	32.3	0	0.0	0	0.0	
1	26	7	6	6	37.5	0	0.0	0	0.0	
1	27	6	5	7	37.6	0.2	0.4	0.3	0.4	
1	28	10	9	7	218.9	1	0.9	1	0.8	
1	29	10	8	22	79.6	0	0.0	0	0.0	
1	30	4	3	7	8.5	0	0.0	0	0.0	
1	31	13	12	1	6.8	0	0.0	0	0.0	
1	32	18	8	24	24.5	1	0.9	0.4	0.2	yes
2	7	6	4	1	0.1	0	0.0	0	0.0	
2	12	5	3	4	5.0	0.9	28.0	0.7	13.5	
2	17	3	1	1	0.3	0.9	30.6	1	30.6	
2	21	5	3	4	8.0	1	5.8	1	3.3	
3	7	9	8	3	6.4	0	0.0	0	0.0	
3	8	9	7	12	46.6	0	0.0	1	1.0	yes
3	10	8	7	2	1.0	0	0.0	0	0.0	
3	15	4	3	1	8.5	0	0.0	0	0.0	
3	17	9	8	1	4.1	0.7	24.5	0.8	24.5	
3	20	11	9	3	5.1	1	17.7	1	15.6	
3	23	9	7	8	57.5	1	2.7	1	2.7	
3	24	6	4	4	2.2	1	5.4	1	3.5	
3	25	6	5	4	5.4	0	0.0	0	0.0	
3	26	6	5	4	4.6	0	0.0	0	0.0	
3	27	23	9	9	148.7	0	0.0	0.2	0.6	yes
3	39	7	5	6	23.0	1	0.3	1	0.1	
3	40	6	5	1	0.8	1	31.5	1	26.8	
3	41	4	3	2	1.0	1	6.7	1	5.1	
3	43	8	6	4	18.3	1	41.7	1	42.9	
17	4	14	13	4	134.7	0	0.0	0	0.0	
17	6	8	5	6	598.3	0	0.0	0	0.0	
17	7	10	8	4	218.2	1	15.5	1	12.4	
17	11	21	12	37	143.9	1	13.4	1	15.5	
17	12	29	17	37	662.6	0.3	0.1	0.5	1.7	
17	14	21	17	9	1,611.3	1	20.5	1	2.2	
17	15	12	10	5	34.3	0.9	4.8	0.8	4.0	
17	17	5	3	6	9.6	0.6	1.9	0.4	0.6	
17	21	4	3	3	7.5	1	35.3	1	30.1	
17	23	19	10	32	566.1	1	12.8	1	5.8	
17	24	20	16	29	970.1	1	0.2	0	0.0	yes
17	26	22	9	36	7,476.3	1	51.1	1	8.7	
17	27	15	7	34	3,675.4	1	58.8	1	32.1	
17	31	11	9	2	20.1	1	1.2	1	4.0	
17	32	28	18	37	338.3	0	0.0	0	0.0	
17	33	21	12	37	500.1	1	0.7	0	0.0	yes
53	25	9	7	13	877.6	0.8	0.3	0.9	0.3	
53	33	5	4	1	1.0	0	0.0	0	0.0	
53	34	8	6	8	304.5	1	8.9	1	0.7	
53	37	2	1	2	5.5	0	0.0	0	0.0	
53	40	5	4	3	11.5	0	0.0	0	0.0	
53	41	4	3	1	1.9	0	0.0	0	0.0	
53	45	2	1	4	6.0	0	0.0	0	0.0	
53	46	4	3	1	0.3	0	0.0	0	0.0	

Table continues on next page.

Table 5 (continued).

Group	Subcategory	$\|N\|$	Min$\|S\|$	Number of stores with $\|S\| < \|N\|$	Total error ($\delta = 0$)	Proportional substitution		Random substitution		Models disagree
						δ^*	% Error reduction[a] (%)	δ^*	% Error reduction[a] (%)	
55	5	8	6	4	359.2	1	26.4	1	22.2	
55	27	9	7	3	189.5	0	0.0	0	0.0	
55	30	8	7	3	1,176.9	0	0.0	0	0.0	
127	2	4	3	5	30.4	0	0.0	0	0.0	
127	3	8	7	11	54.9	0.5	1.5	0.5	1.5	
127	7	6	5	1	5.8	0	0.0	0	0.0	
127	8	10	6	15	52.1	0	0.0	0	0.0	
127	9	7	6	4	6.1	0	0.0	0	0.0	
127	11	7	6	5	20.8	0	0.0	0	0.0	
127	13	9	8	10	49.6	0	0.0	0	0.0	

Note. [a]Error reduction $= 1 - \sum_h \sum_t (\hat{y}_{ht}(\delta) - y_{ht})^2 / \sum_h \sum_t (\hat{y}_{ht}(0) - y_{ht})^2$.

maximize each product's expected revenues minus the cost of disposed inventory.

Some products have case sizes larger than the capacity of a single facing. If only one facing is assigned to such items, the reorder point is negative, no inventory is carried, and sales of the item is zero. Nevertheless, second or third facings can create positive returns. The regular greedy heuristic never assigns the first facing to these products because the marginal benefit of the first facing is zero. Therefore, for products with a case size larger than a facing, the first facing is defined as a combination of as many facings as necessary to fit one case-pack. Also, for perishable products or products with a high demand rate, the gross profit function might be nonconcave: if the inventory level is frequently zero at the end of the period (either because unsold units are disposed of due to perishability or all inventory is sold due to high demand), and the number of cases that fit in the given space does not increase with an assigned facing, the last facing might not be effective. The gross profit function resembles a step function that jumps every time a new case is able to fit into the space defined by the number of facings. This nonconcavity is dealt with by allowing the greedy heuristic to consider allocating one or two facings at a time.

6.4. Recommended Changes and Financial Impact

We applied our estimation and optimization methodology to the data from 37 stores and two categories that we had shelf-space allocation data for. The categories butter & margarine and cookies include 34 subcategories and 234 SKUs. (AP) is solved for each category for a given category of shelf space. The facing allocations for SKUs also determine the space allocated to subcategories. We compare the category gross profit of the recommended assortments with that of the current assortments at Albert Heijn. The gross profit of the recommended system optimized over the peak-load periods is 13.8% higher than that of the current assortment. However, this improvement might not be realized at nonpeak-load periods such as Monday through Wednesday. Weekly gross profits are estimated based on daily demand rates and service levels of peak-load periods for high demand days, and with 100% service level for low demand days. The estimated total improvement in weekly gross profit is 6.2%. Table 6 summarizes these results.

It might be fairer to benchmark our procedure with the assortment Albert Heijn would likely create given recent sales data rather than the existing assortment determined six months ago. The assortment planning principles used at Albert Heijn involve allocating shelf space to products (hence selecting the assortment and reallocating space between subcategories) proportional to their average sales at that store. The continuous numbers are rounded down to discrete facings, and resulting excess space is distributed to the products that had the highest remainders in the round-down procedure. We call this approach the *proportional allocation heuristic*. As seen in Table 6, the proportional allocation heuristic yields a gross profit that is 3.5% higher than the existing assortment at Albert Heijn over the optimized peak-load period and is 0.7% higher in weekly gross profits. The weekly gross profit improvement by the proportional improvement is about one-ninth of the improvement associated with our approach.

According to 2002 financial data, Albert Heijn's gross profit is 25% of its revenues, and its pretax income is 3% of its revenues. Therefore, a 6.2% increase in gross profit is a 1.55% increase in revenues, which increases pretax income to 4.55%. In other words, the impact of our methodology would bring a 52% increase in pretax profits. Based on Albert Heijn's \$10 billion yearly sales, the use of our approach yields a \$155 million increase in pretax profits.

A summary of the recommended changes is presented in Table 7. There were 35 subcategories and 37 stores considered, leading to 1,295 combinations. The recommendation is to reduce the number of products offered in 627 cases,

Table 6. Gross profit improvements by our recommended solution and the proportional allocation heuristic over the current assortments at Albert Heijn.

		% Improvement in peak-load gross profit (%)		% Improvement in weekly gross profit (%)	
Product category		Proportional allocation	Recommended solution	Proportional allocation	Recommended solution
3	Average	3.5	13	1.3	7.0
	Min	−3.1	1.6	−2.7	0.7
	Max	45.8	76.0	15.7	30.9
55	Average	3.6	15	0.3	5.6
	Min	−5.0	0.0	−3.4	0.3
	Max	30.3	65.4	8.8	19.3
Grand average		3.6	13.9	0.7	6.2

Table 7. Summary of recommended changes across all stores and subcategories.

		# of stores with assortment size		Average absolute change in subcategory shelf space (%)	Average absolute change in # facings (%)
Category	Subcategory	Decreased	Increased		
3	7	20	3	15	15
	8	11	6	15	15
	9	10	0	24	24
	10	8	7	15	15
	11	9	0	25	26
	13	37	0	19	21
	14	29	0	28	32
	15	10	1	22	21
	16	6	0	27	27
	17	14	5	25	27
	19	14	3	25	24
	20	17	5	20	20
	21	26	0	30	29
	23	15	5	26	27
	24	15	2	21	22
	25	17	2	22	20
	26	15	1	21	21
	27	2	23	98	93
	39	23	2	28	28
	40	16	3	29	28
	41	27	1	52	52
	42	28	0	23	23
	43	24	2	22	20
	44	26	0	17	17
	47	12	0	21	21
3 Total		431	71	27	27
55	5	35	0	25	23
	6	22	0	23	23
	7	7	0	17	16
	12	33	0	21	21
	15	10	0	31	31
	18	15	0	22	22
	24	27	0	19	18
	27	33	1	11	11
	30	14	2	39	37
55 Total		196	3	23	23
Grand total		627	74	26	26

Table 8. Impact of the substitution rate on assortment size, shelf space, and total number of facings in a subcategory.

	Substitution rate (%)		
	Low	Medium	High
Average magnitude of assortment size increase	77	29	28
Average magnitude of assortment size decrease	−24	−26	−29
Average change in shelf space	12	−3	−3
Average change in number of facings	11	−3	−2

Note. The substitution rate is classified as low if $\delta < 0.3$; high if $\delta > 0.7$; medium otherwise.

add more products in 74 cases, and not change the size of the assortment in the remaining 594 cases. Table 7 also shows that the magnitude of the changes in the shelf space and the number of facings assigned to a subcategory are quite large.

Table 8 illustrates the relation between the magnitude of recommended changes in a subcategory and the substitution rate. Shelf space and facings are transferred from categories with high substitution rates to those with low substitution rates. Similarly, conditional on an increase in assortment size, it seems that the magnitude is larger in subcategories with low substitution rates. These observations support the intuition that broader assortment and higher in-stock rates are more important in categories with higher substitution.

The improvements achieved by our approach stems from three proposed changes to the existing assortment: radical alteration in total space allocated for subcategories, the addition/deletion of SKUs within subcategories, and the facing exchanges between SKUs in subcategories. One of the reasons for facing exchanges between SKUs is due to Theorem 4. Facing allocations have usually been based on sales levels (demand rates). Therefore, a high-demand item always gets more facings. However, maximum inventory levels are usually much higher than the demand rates as a result of positive lead times and case-packs. A high-demand item that has a smaller case-pack does not need several facings to reach its sales plateau. These facings can be assigned to lower demand items with larger case-packs. This benefit could not have been recognized if operational aspects such as case-packs had not been incorporated into gross profit estimation. Table 9 presents a more detailed description of the changes for two examples from cate-

Table 9. Description of the recommended changes to Albert Heijn assortment and breakdown of the gross profit improvement by subcategory for two examples.

	Current solution			Recommended changes			
Subcat.	Space	GP	# SKUs	# SKUs	Space (%)	GP (%)	Other
(a) Store 1,002, category 55, an average example with 12% gross profit improvement.							
5	989	11.824	8	−3	2	35	Added facings to 3 products.
6	715	11.352	3	0	43	1	Added 3 facings to a product.
7	644	8.101	5	0	17	4	Added 1 facing to a product.
12	1,938	33.418	7	0	−29	0	Reduced 1 facing each from 6 products.
15	1,968	22.031	6	0	−41	0	Reduced 1 facing each from 6 products.
18	877	7.661	5	−2	−13	9	Added 1–2 facings to 3 products.
24	1,272	8.691	4	−1	−9	28	Added 1 facing to 1 product, reduced 1 from another.
27	2,870	21.418	9	−3	5	7	Added 1 facing each to 5 products.
30	2,017	30.362	8	−1	50	32	Added 4 facings to a product, 3 to another, 1 to another.
(b) Store 1,141, category 55, an extreme example with 65% gross profit improvement.							
5	1,510	11.75	8	−2	6	67	Added 1 facing to 4 products.
6	1,022	6.164	3	−1	−19	76	Added 2 facings to a product with a very large case size.
7	1,154	6.544	5	0	26	54	Reduced facings from a product with decreasing GP (due to disposals) and increased facings for the others by 1.
12	2,688	37.366	7	−1	−15	6	Dropped a product with negative GP, increased 2 facings for 2 products.
15	2,804	19.03	6	−1	−25	50	Dropped a low-margin product and added 3 facings to a product with higher margin and large case size.
18	1,422	7.721	5	0	48	63	Added facings to 4 products and reduced 1 facing from a product.
24	1,968	1.64	4	−3	−78	−18	Dropped products require 3–4 facings for significant sales because case sizes are twice the capacity of facings, but demand and margins are low.
27	3,978	4.598	9	−5	−26	181	Dropped 4 products with very low GP; added facings to others, all large case size, one also with high margin.
30	3,501	16.081	8	0	79	195	Added 8 facings to 2 products with case sizes 1.5 times facing capacity and 2 facings to 2 products with also large cases.

Note. GP denotes gross profit.

gory 55. The first example (Table 9a) is an average case where the recommended solution achieved a 12% improvement in gross profits. Products with low profit are dropped from the assortment, the number of facings of products with low marginal return are reduced, and the number of facings of those with higher returns are increased. The second example (Table 9b) is from the store that achieved the highest improvement, a 65% increase in profit. Again, most of the changes are facing exchanges between products; particularly, the products with large case-packs are assigned more facings by the algorithm. Subcategories 6, 15, and 27 achieved higher profits with less space. In all three cases, products with low profits are dropped from the assortment, and more facings are assigned to products with large case-packs and larger margins. These benefits underline the importance of considering the profit impact of facing sizes, case-packs, margins, and demand together while developing an assortment plan.

7. Conclusion

The objective of this paper is to develop an algorithmic process to help retailers determine the best assortment for each store. This paper formulates and solves a general and realistic assortment optimization problem. We provide methodologies for estimating the parameters of the model for different sets of available data. Using our optimization algorithm, the retailer can add or delete products from stores that carry less than full assortment and delete products from stores with full assortment. We establish certain structural properties of the assortments suggested by the heuristic optimization algorithm. We find that products with higher demand, higher margin, or smaller physical size should be included first in the assortment and should be assigned more inventory. Products with lower demand variability and smaller case sizes should also be included in the assortment first, but more inventory can be assigned to products with higher demand variance and larger case sizes if the available shelf space is sufficiently high. Finally, we describe the details and the impact of our application at Albert Heijn. We observe the above structural properties in the optimal assortments computed for Albert Heijn as well. Comparing the results of the recommendations of our system with the existing assortments suggests a more than 50% increase in profits. The method described here demonstrates that focusing on the right decisions (the set of products and their stocking levels) and taking operational characteristics of products into account in assortment planning can greatly improve a retailer's profitability.

The assortment optimization model deals with a setting in which shelf space limits total inventory and shelf-space allocation determines the inventory level of each product. We allow for any inventory replenishment system. We consider mature product categories and assume that demand estimates from past data are valid. Our methodology is not suitable for seasonal product categories because we do not deal with a dynamic assortment problem. Still, the optimization model can be used within each season. We believe that our model is applicable to supermarkets, convenience stores, liquor stores, most parts of home improvement stores (e.g., Home Depot), without ruling out other possible retail applications.

Our estimation methodology is also quite general. The specific regression equations will, of course, depend on the context, but any estimation procedure that involves a choice from multiple products can replace the procedure in §4.1. The methodology for estimating choice and substitution simultaneously is general and can be applied to any setting where inventory-transactions data are available. If inventory data are not available, estimating assortment-based substitution requires sales data from multiple stores with varying assortments. Although the model we presented in §4.1 does not lend itself directly to estimating the demand for new products, it is easy to modify our choice models to use the approach suggested by Fader and Hardie (1996). They demonstrate that using product attributes as independent variables in the regression equations rather than using product-specific dummy variables enables the estimation of demand for new products, which can then be included in consideration sets used for assortment optimization.

Acknowledgments

The authors gratefully acknowledge Frank Jansen and Robert van Lunteren of Albert Heijn for their intellectual contributions and support for this project. They thank David Bell, Gerard Cachon, Morris Cohen, Ananth Raman, Paul Zipkin, two anonymous reviewers, and the associate editor for helpful comments.

References

Anderson, S. P., A. de Palma, J. F. Thisse. 1992. *Discrete Choice Theory of Product Differentiation*. MIT Press, Cambridge, MA.

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

Avsar, Z. M., M. Baykal-Gursoy. 2002. Inventory control under substitutable demand: A stochastic game application. *Naval Res. Logist.* **49** 359–375.

Aydin, G., W. H. Hausman. 2003. Supply chain coordination and assortment planning. Working paper, University of Michigan, Ann Arbor, MI.

Ben-Akiva, M., S. R. Lerman. 1985. *Discrete Choice Analysis: Theory and Application to Travel Demand*. MIT Press, Cambridge, MA.

Bretthauer, K. M., B. Shetty. 2002. The nonlinear knapsack problem—Algorithms and applications. *Eur. J. Oper. Res.* **138** 459–472.

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

Bucklin, R. E., S. Gupta. 1992. Brand choice, purchase incidence, and segmentation: An integrated modeling approach. *J. Marketing Res.* **29** 201–215.

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

Campo, K., E. Gijsbrechts, P. Nisol. 2003. The impact of retailer stockouts on whether, how much, and what to buy. *Internat. J. Res. Marketing* **20** 273–286.

Campo, K., E. Gijsbrechts, P. Nisol. 2004. Dynamics in consumer response to product unavailability: Do stock-out reactions signal response to permanent assortment reductions? *J. Bus. Res.* **57** 834–843.

Caprara, A., D. Pisinger, P. Toth. 1999. Exact solution of the quadratic knapsack problem. *INFORMS J. Comput.* **11** 125–137.

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

Cooper, L. G., M. Nakanishi. 1988. *Market-Share Analysis: Evaluating Competitive Marketing Effectiveness*. Kluwer Academic Publishers, Amsterdam, The Netherlands.

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

Dempster, A. P., N. M. Laird, D. B. Rubin. 1977. Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc. B.* **39** 1–38.

Downs, B., R. Metters, J. Semple. 2001. Managing inventory with multiple products, lags in delivery, resource constraints, and lost sales: A mathematical programming approach. *Management Sci.* **47** 464–479.

Dussault, J. P., J. A. Ferland, B. Lemaire. 1986. Convex quadratic programming with one constraint and bounded variables. *Math. Programming* **36** 90–104.

Fader, P. S., B. G. S. Hardie. 1996. Modeling consumer choice among SKUs. *J. Marketing Res.* **33** 442–452.

Fitzsimons, G. J. 2000. Consumer response to stockouts. *J. Consumer Res.* **27** 249–266.

Gallo, G., P. L. Hammer, B. Simeone. 1980. Quadratic knapsack problems. *Math. Programming* **12** 132–149.

Green, P. E., A. M. Krieger. 1985. Models and heuristics for product line selection. *Marketing Sci.* **4**(1) 1–19.

Greene, W. H. 1997. *Econometric Analysis*. Prentice Hall, Englewood Cliffs, NJ.

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

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

Hoch, S. J., E. T. Bradlow, B. Wansink. 1999. The variety of an assortment. *Marketing Sci.* **25** 342–355.

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

Kahn, B. E., D. R. Lehmann. 1991. Modeling choice among assortment. *J. Retailing* **67** 274–275.

Klastorin, T. D. 1990. On a discrete nonlinear and nonseparable knapsack problem. *Oper. Res. Lett.* **9**(4) 233–237.

Kök, A. G. 2003. Management of product variety in retail operations. Ph.D. dissertation, The Wharton School, University of Pennsylvania, Philadelphia, PA.

Kök, A. G., M. L. Fisher, R. Vaidyanathan. 2005. Assortment planning: Review of literature and industry practice. N. Agrawal, S. Smith, eds. *Retail Supply Chain Management*. Kluwer, Amsterdam, The Netherlands.

Kurt Salmon Associates. 1993. Efficient consumer response: Enhancing consumer value in the grocery industry. Food Marketing Institute Report # 9-526, Washington, D.C.

Mahajan, S., G. J. van Ryzin. 1998. Retail inventories and consumer choice. S. Tayur, R. Ganeshan, M. Magazine, eds. *Quantitative Methods in Supply Chain Management.* Kluwer, Amsterdam, The Netherlands.

Mahajan, S., G. van Ryzin. 2001. Stocking retail assortments under dynamic consumer substitution. *Oper. Res.* **49** 334–351.

McFadden, D. 1974. Conditional logit analysis of qualitative choice behavior. P. Zarembka, ed. *Frontiers in Econometrics*. Academic Press, New York.

McGillivray, A. R., E. A. Silver. 1978. Some concepts for inventory control under substitutable demand. *INFOR* **16**(1) 47–63.

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

Nahmias, S., C. P. Schmidt. 1984. An efficient heuristic for the multi-item newsboy model with a single resource constraint. *Naval Res. Logist.* **31** 463–474.

Netessine, S., N. Rudi. 2003. Centralized and competitive inventory models with demand substitution. *Oper. Res.* **51** 329–335.

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

Quelch, J. A., D. Kenny. 1994. Extend profits, not product lines. *Harvard Bus. Rev.* **72** 153–160.

Rajaram, K. 2001. Assortment planning in fashion retailing: Methodology, application and analysis. *Eur. J. Oper. Res.* **129** 186–208.

Rajaram, K., C. S. Tang. 2001. The impact of product substitution on retail merchandising. *Eur. J. Oper. Res.* **135** 582–601.

Schary, P. B., M. Christopher. 1979. Anatomy of a stock out. *J. Retailing* **55** 59–70.

Shugan, S. M. 1989. Product assortment in a triopoly. *Management Sci.* **35** 304–320.

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

Smith, S. A., N. Agrawal. 2000. Management of multi-item retail inventory systems with demand substitution. *Oper. Res.* **48** 50–64.

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

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

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

Wu, C. F. J. 1983. On the convergence properties of the EM algorithm. *Ann. Statist.* **11** 95–103.